Universal Approximation on Path Spaces and Applications in Finance

D I S S E R T A T I O N
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Philipp Schmocker

from

Beatenberg (Bern)

Approved on the application of

Prof. Enrico Giovanni De Giorgi, PhD

and

Prof. Juan-Pablo Ortega Lahuerta, PhD

Dissertation no. 5248

Länggass Druck AG, Bern 2022

Universal Approximation on Path Spaces and Applications in Finance

D I S S E R T A T I O N
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Philipp Schmocker

from

Beatenberg (Bern)

Approved on the application of

Prof. Enrico Giovanni De Giorgi, PhD

and

Prof. Juan-Pablo Ortega Lahuerta, PhD
Prof. Francesco Audrino, PhD

Dissertation no. 5248

Länggass Druck AG, Bern 2022

The University of St.Gallen, School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science, hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St.Gallen, June 13, 2022

The President:

Prof. Dr. Bernhard Ehrenzeller

Extended Dissertation Committee:

Supervisor:	Prof. Dr. Enrico G. De Giorgi
Co-Supervisor:	Prof. Dr. Juan-Pablo Ortega
Programme Committee Member:	Prof. Dr. Francesco Audrino
Additional Member:	Prof. Dr. Christa Cuchiero
Additional Member:	Prof. Dr. Josef Teichmann

Submitted in partial fulfillment of the requirements for the degree of

Ph.D. in Economics and Finance

Abstract

This dissertation studies universal approximation between infinite dimensional spaces, which is of particular interest for applications in Finance. In the finite dimensional case, an interesting class of functions possessing the universal approximation property was discovered eighty years ago: neural networks, which, as concatenation of affine and non-linear functions, are rich enough to approximate every continuous function on a compact set of the Euclidean space arbitrarily well. However, this class of functions only gained wider attention after the turn of the millennium by showing novel and promising applications in the fields of image and speech recognition as well as computer games. Today, neural networks are well known and an integral part of machine learning, a branch of artificial intelligence, used to solve high dimensional and non-linear approximation problems, which were previously inaccessible with conventional methods.

While the universal approximation property of neural networks between finite dimensional spaces was mathematically proven forty years ago, the generalization to infinite dimensional spaces is not straight-forward. Of course, one could argue that everything in our world is finite, such as the number of atoms. Nevertheless, a generalization of neural networks and their approximation property to infinite dimensional spaces would be desirable to solve problems in functional data analysis, dynamical systems and partial differential equations. This would allow to learn path-dependent functionals, such as the payoff of an Asian option, or more general operators, for example the solution operator of a partial differential equation.

In a first application, the universal approximation result between infinite dimensional spaces is used in the framework of Stochastic Portfolio Theory (SPT) to learn an optimal path-dependent portfolio. SPT is a relatively new area of mathematical finance, which attempts to describe equity markets with observable characteristics under realistic assumptions. In this context, a generating function determines the weights of a portfolio and its overall investment behaviour, so that an optimal portfolio can be learned by approximating the optimal generating function. Furthermore, this approach of portfolio gener-

ation has been extended to the path-dependent setting, where the generated portfolio can also depend on the past market trajectory. Since the generating functional now depends on a continuous path, an infinite dimensional object, we use a functional neural network to approximate the optimal path-dependent functional.

In a second application, we use the universal approximation result with infinite dimensional range to approximate the pricing operator of financial derivatives with non-linear dynamics, which corresponds to the solution operator of certain Hamilton-Jacobi-Bellman equations. We use this approach to learn American option prices in a complete market as well as European option prices in an incomplete market. The method is model-free in the sense that it can be applied to any financial market model and the entire value function of the financial derivative of interest can be learned from other known option prices.

In summary, we solve problems in Finance by using machine learning techniques, where we prove the applicability of these methods mathematically through concepts of functional analysis.

Zusammenfassung

Diese Dissertation befasst sich mit universeller Approximation zwischen unendlich dimensionalen Räumen, die für Anwendungen in den Finanzwissenschaften von besonderem Interesse ist. Im endlich dimensionalen Fall wurde eine interessante Klasse von Funktionen mit der universellen Approximationseigenschaft vor achtzig Jahren entdeckt: Es war die Geburtsstunde neuronaler Netzwerke, die als Verkettung affiner und nichtlinearer Funktionen genügend reichhaltig sind, um jede stetige Funktion auf einer kompakten Menge des Euklidischen Raumes beliebig gut zu approximieren. Grössere Aufmerksamkeit erlangte diese Funktionsklasse erst nach der Jahrtausendwende durch neuartige und vielversprechende Anwendungen in der Bild- und Spracherkennung sowie der Computerspiele. Heute sind neuronale Netzwerke ein fester Bestandteil des maschinellen Lernens, einem Zweig der künstlichen Intelligenz, welche zur Lösung hochdimensionaler, nichtlinearer Approximationsprobleme eingesetzt werden, die mit bisherigen Methoden unzugänglich waren.

Während die universelle Approximationseigenschaft von neuronalen Netzwerke zwischen endlich dimensionalen Räumen vor vierzig Jahren mathematisch bewiesen wurde, folgt die Verallgemeinerung auf unendlich dimensionale Räume nicht direkt. Natürlich könnte man argumentieren, dass in unserer Welt alles endlich ist, wie bspw. die Anzahl der Atome. Dennoch wäre eine Ver-allgemeinerung von neuronalen Netzwerken und ihrer Approximationseigenschaft auf unendlich dimensionale Räume wünschenswert, um Probleme in der funktionalen Datenanalyse, dynamischen Systeme und partiellen Differentialgleichungen zu lösen. Dies würde erlauben, pfadabhängige Funktionale, wie die Auszahlung einer asiatischen Option, oder allgemeinere Operatoren, bspw. den Lösungsoperator einer partiellen Differentialgleichung, zu lernen.

In einer ersten Anwendung wird die universelle Approximationsresultat zwischen unendlich dimensionalen Räumen im Rahmen der Stochastischen Portfoliotheorie (SPT) benutzt, um ein optimales pfadabhängiges Portfolio zu lernen. SPT ist ein relativ neues Gebiet der Finanzmathematik, das versucht, Aktienmärkte mit beobachtbaren Eigenschaften unter realistischen Annahmen zu beschreiben. In diesem Zusammenhang bestimmt eine generierende Funk-

tion die Gewichte eines Portfolios und sein gesamtes Investitionsverhalten, so dass ein optimales Portfolio durch Approximation der optimalen generierenden Funktion gelernt werden kann. Darüber hinaus wurde dieser Ansatz der Portfolio-Generierung ins pfadabhängige Setting erweitert, so dass das generierte Portfolio auch von der Vergangenheit des Markts abhängen kann. Da das generierende Funktional nun von einem unendlich dimensionalen Pfad abhängt, verwenden wir ein funktionales neuronales Netzwerk, um das optimale pfadabhängige Funktional zu approximieren.

In einer zweiten Anwendung benutzen wir das universelle Approximationsresultat mit unendlich dimensionalen Wertebereich, um den Preisbildungsoperator von Finanzderivaten mit nichtlinearen Dynamiken zu approximieren, der dem Lösungsoperator bestimmter Hamilton-Jacobi-Bellman-Gleichungen entspricht. Wir verwenden diesen Ansatz, um amerikanische Optionspreise in einem vollständigen Markt sowie europäische Optionspreise in einem unvollständigen Markt zu lernen. Die Methode ist modellfrei in dem Sinne, dass sie für ein beliebiges Finanzmarktmodell angewendet werden kann, und die gesamte Wertefunktion des interessierenden Finanzderivat aus anderen Optionspreisen, die im Markt bekannt sind, gelernt werden kann.

Zusammenfassend lösen wir Probleme in den Finanzwissenschaften durch maschinelles Lernen, wobei wir diese Methoden mathematisch durch Konzepte der Funktionalanalysis beweisen.

Acknowledgements

A man provided with paper, pencil, and rubber, and subject to strict discipline, is in effect a universal machine – Alan M. Turing

First, I would like to thank my thesis supervisor **Prof. Dr. Enrico G. De Giorgi** for accepting me as PhD student. In addition, I would like to thank **Prof. Dr. Francesco Audrino** for representing the program committee.

My warmest thanks go to **Prof. Dr. Christa Cuchiero**, who supervised me during my doctoral studies with all her kindness and offered me her help at any time. I am very impressed by her broad research knowledge, which in combination with her exactness could not have been a better supervision for me. Even from her vacations in Estonia, she took time for me and provided me with her helpful feedback on the draft of my dissertation to complete the research projects. Her ideas and motivating support have contributed significantly to the completion of this thesis. I am endlessly grateful to her for her motivating support, her kind supervision and for the time she has always taken for me – thank you very much, Christa!

Equally, these warmest thanks go to **Prof. Dr. Josef Teichmann** for supervising me during my doctoral studies in his very amicable way. He challenged me not only in a mathematical way, but also physically with jogging rounds in Zurich or climbing the Mount Pilatus with his research group. I am very impressed by his endless enthusiasm for research, which inspired me every day and always fueled my enjoyment for problems in the fields of machine learning and mathematical finance. His extensive knowledge goes far beyond these topics and I have always enjoyed talking to him about politics and problems in society. I am more than grateful for his motivating support and having the possibility to work with him – thank you so much, Josef!

Moreover, I would particularly like to express my gratitude to **Prof. Dr. Ariel Neufeld** for inviting me as an exchange student to his research group at the Nanyang Technological University (NTU) in Singapore. With his refreshing and positively energetic manner, he supported me not only during the exchange but also until the completion of this dissertation. I benefited a lot from his ideas

and knowledge in the areas of mathematical finance and stochastic processes. For me, it was a great experience to work with him in Singapore, even during the Singaporean circuit breaker – thank you very much, Ariel!

In addition, I would like to thank **Prof. Dr. Juan-Pablo Ortega** for his motivating support during my doctoral studies and being the co-supervisor of this thesis. I am very grateful for his open-minded way, for always listening to me and for our inspiring conversations around the world, in St. Gallen, Zurich, Constance and Singapore – many thanks also to you, Juan-Pablo!

Furthermore, I would like to thank all my colleagues at the University of St.Gallen, where I met Andrin, Anna-Lena, Cédric, Chen, Daniele, Dominik, Giovanna, Jan, Jeannine, Lukas, Matthias, Robert, Simon, Thomas, and Tri, and of course our sunshine Margit. Moreover, I would like to thank the research group at ETH Zurich and all its members of the Friday evening seminars and the subsequent dinners, as well as the research group in Singapore for their welcoming manner. In addition, I am very grateful to my lovely couple, Melissa and Thapelo, for making me feel home in Singapore and keeping up the good spirit during the circuit breaker, as well as Elli and Julian for the shared adventures on the small island.

Finally, I want to express my profound gratitude to my parents Peter and Karin for their unconditional love and unfailing support over the past years. You continuously encouraged me in proceeding my studies and supported me throughout the process of doing research and writing this thesis.

Many thanks go to my twin brother Oliver and his fiancée Laura for the social activities during the writing of this thesis and for always cheering me up. Last but not least, I would like to thank all my friends for their support and unforgettable moments we shared around the world.

This accomplishment would not have been possible without everyone of you. My warmest thanks to all of you!

Singapore, July 5th, 2022 Philipp Schmocker

List of Mathematical Symbols

General Symbols

$\mathbb{N} = \{0, 1, ...\}$	Set of natural numbers
$\mathbb{N}^d$	Set of multi-indices $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}^d$
$\mathbb{N}_k^d$	Set of multi-indices $\alpha \in \mathbb{N}^d$ with $\lvert\alpha\rvert \leqslant k$
$\mathbb{R}$, $\mathbb{R}^d$	Set of real numbers and Euclidean space
(X, d_X), $(X, \lVert\cdot\rVert_X)$, (X, τ_X)	Metric space, normed space, topological space
$\mathcal{P}$	Family of seminorms $p : X \to \mathbb{R}$
$L(X;Y)$, $X^* := L(X, \mathbb{R})$	Space of cont., linear functions (with $Y = \mathbb{R}$)
$C^0(X;Y)$, $C^0(X)$	Space of continuous functions ($Y = \mathbb{R}$)
$C_b^0(X;Y)$, $C_b^0(X)$	Space of bounded, cont. functions ($Y = \mathbb{R}$)
$C^k(X;Y)$, $C^k(X)$	Space of k-times db. functions ($Y = \mathbb{R}$)
$C_b^k(X;Y)$, $C_b^k(X)$	Space of bded. and k-times db. functions ($Y = \mathbb{R}$)
$C^\infty(X;Y)$, $C^\infty(X)$	Space of smooth functions ($Y = \mathbb{R}$)
$C_c^\infty(X;Y)$, $C_c^\infty(X)$	Space of comp. supp., smooth functions ($Y = \mathbb{R}$)
$f'(x)$, $f^{(k)}(x)$	First and kth derivative of a function $f : \mathbb{R} \to \mathbb{R}$
$\partial^\alpha f(x)$	Partial derivative of $f : \mathbb{R}^d \to \mathbb{R}$ of degree $\lvert\alpha\rvert$

Universal Approximation

$\psi : X \to]0, \infty[$	Weight function
$\varphi : X \to Y$,	Abstract neural network
$\mathcal{NN}_{d,M}^{\rho}$	Set of conventional neural networks
$\mathcal{FN}_{d,M}^{\rho}$	Set of functional neural networks
$\mathcal{H}$, $\mathcal{R}$, $\mathcal{L}$	Additive family, activation family, linear readout
$\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$	Set of abstract neural networks with $(\mathcal{H}, \mathcal{R}, \mathcal{L})$
$\mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$	Set of randomized ANNs with $(\mathcal{H}, \mathcal{R}, \mathcal{L})$
τ_c^0, τ_c^k	Compact convergence topology (of order k)
τ_ψ^0, τ_ψ^k	Weighted topology (of order k)
$\mathcal{B}_\psi(X;Y)$, $\mathcal{B}_\psi(X)$	Weighted space of cont. functions ($Y = \mathbb{R}$)
$\mathcal{L}\mathrm{ip}^k(X;Y)$, $\mathcal{L}\mathrm{ip}^k(X)$	Space of $\mathcal{L}\mathrm{ip}^k$-functions ($Y = \mathbb{R}$)
$\mathcal{L}\mathrm{ip}_\psi^k(X;Y)$, $\mathcal{L}\mathrm{ip}_\psi^k(X)$	Weighted space of $\mathcal{L}\mathrm{ip}^k$-functions ($Y = \mathbb{R}$)
$(U_i, \phi_i)_{i \in I}$	Coordinate charts (atlas) of a manifold
$\mathcal{A}$, $\mathcal{W}$	(Polynomial) Subalgebra and submodule

$\mathrm{Lip}(S;Y)$, $\mathrm{Lip}_0(S;Y)$	Sp. of Lipschitz functions (preserving the origin)
$(Æ(S), \|\cdot\|_{Æ(S)})$	Arens-Eells space on (S, d_S)
$G\Omega_\alpha([0,T];\mathbb{R}^d)$	Space of geometric α-Hölder rough paths

Learning Arbitrages in SPT

Δ^d, $\overline{\Delta}^d$	Open and closed simplex
(Λ_T, d_∞), $(\Lambda_T^\alpha, d_\infty)$	(Weighted) Sp. of stopped (α-Hölder cont.) paths
$F : \Lambda_T \to \mathbb{R}$	Non-anticipative functional
$\mathcal{D}$, ∇	Horizontal and vertical derivative
$\mathbb{C}^{0,0}(\Lambda_T)$, $\mathbb{C}^{p,q}(\Lambda_T)$	Sp. of cont. (smooth) non-anticipative functionals
$S : \Delta^d \to [0,\infty[$	Generating function
$F : \Lambda_T \to [0,\infty[$	Generating functional
μ, Z_μ	Market portfolio and its wealth
π, Z_π	Portfolio and its wealth
$(\mathcal{FG}, d_{\mathcal{FG}})$	Space of functionally generated portfolios

Option Pricing with ANNs

$X = (X_1(t), ..., X_d(t))_t$	Price process of risky assets
$X_0 = (X_0(t))_t$	Additional bank account
$\widetilde{X} = (\widetilde{X}_1(t), ..., \widetilde{X}_d(t))_t$	Price process of discounted risky assets
$(D_{st})_{s\leqslant t}$	Discount factor
$\mathcal{A}$	Infinitesimal generator
θ, V_θ	Trading strategy and its value process
Θ_{adm}	Space of admissible trading strategies
$\mathbb{Q} \sim \mathbb{P}$	Equivalent martingale measure
$C \in L^0(\mathcal{F}_T)$	$\mathcal{F}_T$-measurable financial option
$H^k(E)$, $W^{k,p}(E)$	Sobolev spaces of functions $f : E \to \mathbb{R}$
$u(t,x)$, $g : E \to \mathbb{R}$	Value function of option with payoff $g : E \to \mathbb{R}$
$H(t,x,z,p,r)$	Hamiltonian
$(T_{s,t})_{t\leqslant s}$	Semigroup of partial differential equation (PDE)
$\mathcal{U}_{t,T}$	Space of control pr. $(c(s))_{s\in[t,T]}$ with values in U
$\mathcal{T}_{t,T}$	Set of stopping times τ with values in $[t,T]$

Contents

1 Introduction **1**
1.1 Motivation and Terminology 1
1.1.1 Terminology in Machine Learning 1
1.1.2 Impact of Machine Learning in Economics and Finance 3
1.1.3 Significance of the Thesis for Economics and Finance . . 6
1.2 Mathematical Preliminaries 8
1.2.1 Metric Spaces and Banach Spaces 8
1.2.2 Topological Spaces and Convenient Vector Spaces . . . 16
1.2.3 Stone-Weierstrass and Nachbin's Theorem 22
1.2.4 Stochastic Analysis and Pathwise Ito Calculus 26

I Universal Approximation **29**

2 Universal Approximation in Finite Dimension **31**
2.1 Neural Networks 31
2.2 Universal Approximation 33
2.3 Universal Approximation with Derivatives 35
2.4 Recurrent Neural Networks 38

3 Universal Approximation on Weighted Space **41**
3.1 Motivational Examples between Infinite Dimensional Spaces . . 42
3.2 Weighted Space and Stone-Weierstrass 44
3.2.1 Functions on Weighted Space 44
3.2.2 Weighted Stone-Weierstrass Theorem 48
3.3 Neural Networks in Infinite Dimension 54
3.3.1 Abstract Neural Networks 54
3.3.2 More Examples of Abstract Neural Networks 58
3.3.3 Universal Approximation of Abstract Neural Networks . 61
3.4 Smooth Neural Networks in Infinite Dimension 68
3.4.1 Differentiable Lipschitz Functions on Weighted Space . 68

3.4.2 Weighted Nachbin Theorem with Convenient Calculus . 70
3.4.3 Universal Approximation of Smooth Abstract Neural Networks 84
3.5 Randomized Abstract Neural Networks 88
3.6 Lipschitz Functions and Signature Methods 93
3.6.1 Predual of Lipschitz Functions 93
3.6.2 Signature of Geometric Rough Paths 97

4 Universal Approximation on Space of Stopped Paths 103
4.1 Non-Anticipative Functional Calculus 103
4.2 Functional Neural Networks 105
4.3 Universal Approximation with Derivatives 107
4.4 Examples of Learned Non-Anticipative Functionals 112

II Learning Arbitrages in Stochastic Portfolio Theory 119

5 Stochastic Portfolio Theory 121
5.1 Market and Portfolios 121
5.2 Market Behaviour and Diversity 127
5.3 Functionally Generated Portfolios 129
5.4 Path-Dependent Setting 133

6 Pathwise Stochastic Portfolio Theory 137
6.1 Market Weights and Portfolios 137
6.2 Functionally Generated Portfolios 141
6.3 Path-Dependent Setting 144

7 Learning an Optimal Portfolio 147
7.1 Approximation of Functionally Generated Portfolio 147
7.2 Performance Functional 151
7.3 Path-Dependent Setting 156
7.4 Numerical Examples 157
7.4.1 Recovering Generating Function 158
7.4.2 Outperforming the S&P500 162

III Option Pricing with Abstract Neural Networks 167

8 Option Pricing in a Financial Market 169
8.1 Market Model 169
8.2 Sobolev Spaces 175
8.3 Value Function of Financial Option 178

9 Hamilton-Jacobi-Bellman Equation **183**
9.1 Hamiltonian and Comparison Principle 183
9.2 Stochastic Optimal Control 187
9.3 Pricing Equation of American Option 191
9.4 Super-Replication Price in Stochastic Volatility Model 192

10 Abstract Neural Network and Numerical Examples **195**
10.1 Approximation of the Solution Operator 196
10.2 Learning American Options 203
10.3 Learning Super-Replication Price with Stochastic Volatility . . 204

Chapter 1

Introduction

The aim of this thesis is to solve problems in Finance using novel techniques from the field of machine learning, which were previously inaccessible with conventional methods. We first give an introduction into the terminology of machine learning and motivate the topic of the thesis from an economic point of view, followed by the mathematical notation and preliminaries, which are later used in this dissertation.

1.1 Motivation and Terminology

Thanks to the breakthrough of machine learning technologies after the turn of the millennium in the areas of image and speech recognition as well as computer games, machine learning has become an integral part of our daily lives. In the following, we will go into more details regarding the different terms used in machine learning and motivate the various applications in the world of Economics and Finance.

1.1.1 Terminology in Machine Learning

The term machine learning appeared the first time in the technical report [Sam59] of Arthur L. Samuel in 1959 written for IBM. However, it took until the turn of the millennium when Tom M. Mitchell came up with a more formal definition for algorithms in machine learning: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (see [Mit97]). This definition is similar to the rather operational definition of machines from Alan M. Turing in [Tur50], where the question "Can machines think?" is now replaced by "Can machines do what we (as thinking entities) can do?" (see [Har08]).

In the current literature, the terms machine learning and artificial intelligence are often abusively interchanged, creating confusion among practitioners and non-experts, which is why it is important to distinguish between the different terminologies. Machine learning has actually emerged from the quest for artificial intelligence, where we refer to the textbooks [Mit97] and [RN16] for more background. Meanwhile, many new terms have been established in the field of machine learning, which is why we give a brief overview of the most important terminologies below and refer to [KP98] for a complete glossary.

Artificial intelligence is an umbrella term for the others, describing general machines trying to solve problems, which, unlike humans and animals, lack consciousness and emotional feelings. *Data science*, in turn, is an interdisciplinary subfield of artificial intelligence, which uses scientific methods and computer algorithms to extract knowledge from structured and unstructured data. Finally, *machine learning* is a particular branch of data science, which studies the computer algorithms improving automatically through experience. Besides some other methods, these algorithms can be generally classified into three groups: supervised learning, unsupervised learning and reinforcement learning, see [Mit97] and [RN16].

In *supervised learning*, the computer is typically provided with a set of input data (features) and output data (targets), for which a model should be estimated to predict other outputs. Hereby, the term prediction does not only incorporate forecasting of a time series, for example, but also applies to a setting with labelled observations, where the model is first learned on a given training set (input and output data set), and one would like to predict the target depending on an independent test set. This machine learning method does not only solve the estimation problem, but tries to achieve a goodness of fit in an independent test set by minimizing a given loss function. Typical tools for supervised learning are regularized regression (lasso/ridge regression and elastic net), random forest, regression trees, support vector machines (SVMs), and neural networks (NNs). The difference of these machine learning methods compared to other traditional econometric and statistical approaches is the data-driven model selection, which means that the model is determined depending on the given training data rather than just performing a single estimation.

In *unsupervised learning*, the computer finds clusters of observations which are similar in terms of the input data, where the algorithm learns the objective only based on the input data and should detect hidden patterns by itself. This method is commonly used in video, image and speech recognition, for which there exists a variety of different tools such as hierarchical clustering, k-means clustering, mixture models, method of moments, autoencoders, and generative adversial networks (GANs). The typical output in unsupervised machine learning is a partition of the input data, where the elements in each partition are

similar to each other according to a given metric. These tools are very useful as a data-driven way to find similar research articles or news items, which can then be added as additional variable to another given model.

In *reinforcement learning*, the computer interacts with the environment and tries to maximize a certain objective function (reward function), and thus differs from supervised learning by not requiring labelled data as training set. Typical examples can be found in the areas of game theory, control theory, information theory and multi-agent systems, which have applications in our daily lives, such as robots and self-driving cars. Besides these three main areas, there are also methods in *dimension reduction* and *meta learning*.

In all these methods, there is usually a trade-off between the expressiveness of the model, i.e. the complexity of the model, and the risk of overfitting, i.e. when the model becomes too flexible compared to the sample size. In the latter case, the model is learned relatively well on the training set, but its generalization ability on an independent test set is rather poor. There are many tools to avoid these two extreme phenomena, e.g. cross-validation, regularization or averaging over many different models.

These novel machine learning techniques have nowadays different applications ranging from Biology, over Computer Science, Economics, Finance, Marketing, Robotics to Sport Sciences, whereas one should definitely not forget ethical issues regarding the learned bias of cultural prejudices.

1.1.2 Impact of Machine Learning in Economics and Finance

Various leading economists predict that machine learning will have a groundbreaking impact on Economics and Finance (see, e.g., [Ath19] for Economics, [SS18] for Business and Marketing, [EKOO19] for Quantitative Finance, and [AD19] for Risk Management). The first attempts of machine learning applications in Economics or Finance took place in the 1980s, with, for example, the premature artificial intelligence model of Michael D. Cohen and Robert Axelrod in [CA84] from the year 1984, where a checker player learns based on his past experiences during the game, which considerably outperformed other state-of-the-art models at this time. Another interesting application was discovered by Halbert White in [Whi88], where neural networks have been used to find non-linear regularities in the price movements of the IBM stock data in order to predict future returns. However, the most important and successful applications of machine learning in Economics and Finance were only developed after the turn of the millennium, of which we highlight some examples at the end of this subsection, focusing in particular on the applications in Quantitative Finance to which this dissertation contributes.

So far, much of the empirical work in Economics and Finance consisted of

choosing a suitable regression model and estimating its parameters on the full data set, which relies on statistical theory in order to obtain predictions and confidence intervals. Since the primary interest of this empirical work is only focused on the causal effect, machine learning methods bring a new, data-driven alternative by viewing the estimation procedure as algorithms, where many different models are compared to each other. While the trade-off between expressiveness and overfitting relied in traditional approaches of Econometrics on consistency and asymptotic normality, machine learning approaches are not concerned with this kind of parameter uncertainty anymore and use the performance evaluation on an independent test set. In addition, classical models in Econometrics and Statistics needed an identification assumption (that the parameter can be identified with infinitely many data), whereas machine learning does not address this concern at all and makes the model credible with the evaluation on an independent test set. These two differences make machine learning methods better applicable, also to more complicated problems in Economics and Finance.

Another reason for the success of machine learning methods in Economics and Finance is the so-called technique of "Stochastic Gradient Descent (SGD)", which is used for the estimation procedure of many models, such as neural networks. SGD is a method to optimize a complicated objective (loss) function with respect to the model parameters by averaging over the gradient evaluated at different sample data. This estimate of the gradient is noisy, but unbiased, and follows the idea that it is more efficient to descend into a valley by taking many steps in various directions instead of taking fewer steps in the right direction. In addition, SGD prevents from getting stuck in a local minima, where the model is good but not optimal, and it can lead to performance improvements, which allows the estimation of more complex models.

In addition, the digitalization across the global world is leading to the creation of many economic and financial data sets, which can be fed into machine learning algorithms. These data sets could range from individual consumer behaviour to explicit financial transaction data, which are both recorded at high time frequency. Machine learning methods are able to process these large amounts of data by extracting the relevant information and thus work like an automatic research assistant, only faster and more efficiently. Moreover, the performance of machine learning methods will improve in the next years due to new innovations and faster computers, which is a consequence of the rapidly evolving technology, where anyone can easily implement algorithms with the open-source software `Python`, and where the code of many excellent machine learning algorithms are publicly available. Therefore, machine learning methods will certainly continue to have a major impact on Economics and Finance. In the following, we will now give some examples of machine learning applications for each field Economics, Business and Marketing, Finance as well as Risk

Management, which already showed promising results when comparing them to traditional methods.

For *Economics*, [Ath19] predicts that machine learning will lead to a significant change in many economic fields. For example, the measurement and analysis of an economic variable and the result of an economic outcome can be made at a more granular level and over a longer time period. Moreover, large-scale administrative data sets as well as documentation and storage might be easier to handle at working places, which simplifies interdisciplinary work and will change the organization of a company. Machine learning will also affect the interplay between the government, industry, education, and health, where for example teaching might be changed more towards the topics of computer science and data analytics skills. However, there are still some open problems regarding policy issues such as insurance, liability, safety, data sharing, fariness, and many others, which need to be addressed for further applications in Economics.

For *Business and Marketing*, the authors of [SS18] predict that machine learning brings the fourth industrial revolution in the next few decades. In sales, the machine learning techniques automate the standard selling activities by taking into account the recorded customer behaviour. Moreover, machine learning models can simulate buying centers with all their complexities, informing sales people with all the respective data, which can be used to anticipate demand increase or potential pitfalls.

For *Quantitative Finance*, the authors of [EKOO19] note that the recent advances in machine learning are already finding commercial applications in the financial industry. Indeed, these tools are perfectly applicable to the financial industry with large available data sets such as stock market prices and transaction data. For example, the everyday task in a bank of calibrating a financial market model with historical data was driven so far by the model's computational tractability rather than matching the observed data. For this purpose, neural network approaches offer a complete new perspective, such as in [Her17] and [BS18], as well as in [HMT21] by including rough volatility, and additionally in [CKT20] for a general local volatility model with GANs. On the other hand, banks offer and sell financial products such as options and other derivatives, for which the price of such a contract and the corresponding hedging strategy needs to be computed. Because pricing and hedging are in general very non-linear (in incomplete markets), machine learning techniques offer a new approach and outperform the classical methods. For example, semi-reccurent neural networks were successfully applied in [BGTW19] for pricing and hedging financial derivatives, whereas [LNP19] showed that rough paths can be used to learn the option payoff and the corresponding hedging strategy. Moreover, [BCJ19] applied machine learning techniques to optimal stopping, which can be used for American option pricing, while [GHJW18] approximated

the solution of the associated Black-Scholes partial differential equation of an European option. Another example is the general portfolio selection problem, which can be approached with novel machine learning techniques such as the problem of maximizing utility from terminal wealth, see for example [FS03] with a probabilistic model or [GTX19] with reinforcement learning. Within the framework of Stochastic Portfolio Theory (SPT) initiated by [Fer02], the application of Gaussian processes in [KSV16] showed promising results to obtain optimal portfolios. Another common machine learning application consists of stock return predictions, mostly by using neural networks, for which [Son05] gives a useful overview. Hereby, news sentiment and attention variables obtained from social media and news article can be used as additional characteristics to improve the return prediction, which was applied in [ASB20] for the prediction of stock market volatility. On the other hand, dynamic programming is used for algorithmic trading in [CJK16], where the traders learn from changes in prices and update their predictions. This dissertation contributes to this field of Quantitative Finance by proposing additional interesting machine learning applications.

For *Risk Management*, [AD19] describes how machine learning techniques will transform the fields of credit risk, market risk, operational risk, and compliance, by also mentioning the practical limitations of these novel methods because of the lack of suitable data management policies and necessary skillsets in firms. Hereby, machine learning methods outperform traditional models because the data preparation allows for partially incomplete data and the techniques allow for a better semantic understanding of unstructured data. In addition, the identification of various risks can be recognised by machine learning through hidden patterns, and the methods can handle the huge amount of unconventional data in financial risk management.

Altogether, machine learning will have a substantial impact on the fields of Economics and Finance by providing new methods, which outperform the classical approaches by the nature of these new algorithms.

1.1.3 Significance of the Thesis for Economics and Finance

Overall, machine learning techniques can be used in Quantitative Finance in a very broad range to achieve different goals, for which we have mentioned some particular examples in the previous section. This dissertation contributes twofold to the steadily growing area of machine learning applications in Quantitative Finance. On the one hand, we provide a theoretical framework for a better understanding why machine learning methods work so well by relying on mathematical concepts borrowed from Functional Analysis, Topology, Probability Theory and Statistics. On the other hand, we apply these theoretical

results to develop new machine learning methods, which are able to solve different non-linear problems in Quantitative Finance such as finding a portfolio outperforming the S&P500 or pricing a complicated financial contract.

The first theoretical part of this thesis introduces the generalization of neural networks to high and infinite dimensional spaces, which can be seen as contribution to the area of supervised machine learning. Artificial neural networks were discovered by Warren McCulloch and Walter Pitts in [MP43] and are inspired by the functionality of a human brain, where the input data is fed into the system, passes through various connections (called synapses in Biology) and so-called neurons, with the information being transformed in the latter, and is finally returned on the other end as output. This system can be described in mathematical terms by the concatenation of affine and non-linear maps, where the affine maps represent the connections between the different neurons and where the non-linear map is applied to transform the information of the data. In an economic setting, such neural networks can therefore be seen as a model representing a certain functional relation, for which the so-called universal approximation theorems proved by [Cyb89] and [Hor91] ensure under some assumption the existence of such an approximating neural network.

The infinite dimensional generalization of these neural networks, which we shall call *abstract neural networks (ANN)* in the following, are able to process mathematically more abstract data (infinite dimensional data) such as continuous paths, discontinuous functions or (probability) measures. If both the input and output spaces are finite dimensional Euclidean spaces, we speak of learning a function $f : \mathbb{R}^n \to \mathbb{R}^d$, while for an infinite dimensional input space we would learn a functional $F : X \to \mathbb{R}$, where X could be the space of continuous paths. The third case is learning an operator $T : X \to Y$, where both the input and output spaces are infinite dimensional, e.g. a continuous path is transformed into another one. Since economic or financial data are often of this functional type (see [RS05]), we approximate an economic relation with an abstract neural network. Therefore, this thesis contributes as a theoretical framework to overcome the black box image of neural networks and shows in universal approximation results that ANNs be used to approximate such relations represented by a function, a functional or an operator, where not only the value of the function(al)/operator can be approximated but also the (functional) derivatives, which corresponds to the change of the value of the function(al)/operator compared to the change in the input data.

The applications of these universal approximation results for abstract neural networks between infinite dimensional spaces have various and far-reaching applications, in particular in the areas of Economics and Finance. In many cases, economic data are mathematically more abstract (i.e. infinite dimensional) and given by paths (e.g. historical gross domestic product) or surfaces (e.g. implied volatility surfaces), which can be used as input data of the abstract

neural network, whereas the output data might again be infinite dimensional such as another path (e.g. economic growth) or another function (e.g. trading strategy). Moreover, a contract in Finance can depend on the whole path of the stock price such as it is the case for an Asian option, defined as the average over the historical path, where the stock price might also have jumps. These so-called path space functionals are in particular interesting for pricing and hedging financial derivatives, where the payoff depends on the entire history of the stock price. All these cases can be covered by the theoretical framework of this dissertation, where we provide besides randomized ANNs and a rough path approximation result two particular applications in the second and third part related to optimal portfolios in Stochastic Portfolio Theory (SPT) and general option pricing in Mathematical Finance.

Of course, there are also particular cases which are not covered by the theoretical framework of this thesis, in particular when the output space admits a more complicated structure than a vector space. This is for example the case for normed probability measures as output space, where the sum of two probability measure is no longer a probability measure anymore, or more generally for geometric learning on more complicated manifolds. For this purpose, [KB20], [KP21] and [Kra21] provide useful results to approximate an infinite dimensional neural network with output space given by a manifold.

Altogether, this dissertation establishes a comprehensive theoretical framework for various machine learning applications with mathematically proven results using concepts from Functional Analysis, which can be directly applied to problems in the world of Economics and Finance.

1.2 Mathematical Preliminaries

In this section, we introduce the mathematical notation and the main preliminaries in form of already existing mathematical results in the literature. We start with examples of metric spaces and Banach spaces, extend the considerations to more general topological (vector) spaces, state several existing Stone-Weierstrass theorems, and introduce the basic foundations of Stochastic Analysis.

1.2.1 Metric Spaces and Banach Spaces

As usual, $\mathbb{N} = \{0, 1, 2, ...\}$ denotes the set of natural numbers, whereas $\mathbb{Z} = \{..., -1, 0, 1, ...\}$ represents the set of integers. In addition, $\mathbb{R}$ denotes the set of real numbers and we define the d-dimensional vector space $\mathbb{R}^d$ over $\mathbb{R}$ as vectors

$x = (x_1, ..., x_d)^\top$, with $x_i \in \mathbb{R}$, which is equipped with the usual Euclidean norm

$$\|x\| = \left(\sum_{i=1}^{d} |x_i|^2 \right)^{\frac{1}{2}}, \tag{1.2.1}$$

for $x \in \mathbb{R}^d$. More general, X is called a *vector space (over $\mathbb{R}$)* if the vector space axioms are satisfied for the vector addition $X \times X \ni (x, y) \mapsto x + y \in X$ and scalar multiplication $\mathbb{R} \times X \ni (\lambda, x) \mapsto \lambda x \in X$. We now introduce introduce metric spaces and Banach spaces.

Definition 1.1. *For a set X, the map $d_X : X \times X \to \mathbb{R}$ is called a* metric (on X) *if the following three axioms are satisfied:*

(i) Positive Definiteness: $d_X(x, y) \geqslant 0$, *for all $x, y \in X$, and $d_X(x, y) = 0$ if and only if $x = y$,*

(ii) Symmetry: $d_X(x, y) = d_X(y, x)$, *for all $x, y \in X$,*

(iii) Triangle Inequality: $d_X(x, z) \leqslant d_X(x, y) + d_X(y, z)$, *for all $x, y, z \in X$.*

If $d_X : X \times X \to \mathbb{R}$ is a metric on X, the pair (X, d_X) is called a metric space.

Example 1.2. *The real numbers $\mathbb{R}$ with distance function $d(x, y) = |x - y|$ form a metric space. More general, the Euclidean space $\mathbb{R}^d$ equipped with the Euclidean metric $d(x, y) = \|x - y\|$ is a metric space, where $\|\cdot\|$ denotes the Euclidean norm in* (1.2.1).

Definition 1.3. *For a vector space, the map $\|\cdot\|_X : X \to \mathbb{R}$ is called a* norm (on X) *if the following three axioms are satisfied:*

(i) Positive Definiteness: $\|x\|_X \geqslant 0$, *for all $x \in X$, and $\|x\|_X = 0$ if and only if $x = 0$.*

(ii) Absolute Homogeneity: $\|\lambda x\|_X = |\lambda| \|x\|_X$, *for all $\lambda \in \mathbb{R}$ and $x \in X$.*

(iii) Triangle Inequality: $\|x + y\|_X \leqslant \|x\|_X + \|y\|_X$, *for all $x, y \in X$.*

If $\|\cdot\|_X : X \to \mathbb{R}$ is a norm on X, the pair $(X, \|\cdot\|_X)$ is called a normed (vector) space.

Remark 1.4. *Every normed vector space $(X, \|\cdot\|_X)$ can be seen as a metric space by defining $d_X(x,y) = \|x-y\|$, but the reverse is not necessarily true. For example, the trivial metric on $\mathbb{R}$ defined by*

$$d(x,y) = \mathbb{1}_{\{x=y\}} := \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

is not induced by a norm (except for the zero space $X = \{0\}$) since it violates the absolute homogeneity condition in Definition 1.3(ii). However, if X is a vector space and the metric $d_X : X \times X \to \mathbb{R}$ is absolutely homogeneous and translational invariant, i.e. $d_X(\lambda x, \lambda y) = |\lambda| d_X(x,y)$ and $d_X(x+z, y+z) = d_X(x,y)$, for all $\lambda \in \mathbb{R}$ and $x, y, z \in \mathbb{R}$, then $\|x\|_X = d_X(x,0)$ satisfies the norm axioms.

Example 1.5. *The following are examples of metric and normed vector spaces.*

(i) Euclidean Space: *The space $\mathbb{R}^d$ together with metric $d(x,y) = \|x-y\|$ is a metric space, and forms together with the norm $\|\cdot\|$ defined in* (1.2.1) *a normed vector space.*

(ii) Space of Sequences: *For some $1 \leqslant p < \infty$, let l^p be the space of sequences $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ whose l^p-norm*

$$\|x\|_{l^p} = \left(\sum_{n\in\mathbb{N}} |x_n|^p \right)^{\frac{1}{p}}$$

is finite. Then, $(l^p, \|\cdot\|_{l^p})$ is a normed vector space and for $p \leqslant q$, the embedding $l^q \subseteq l^p$ holds true by the Hölder inequality (see Lemma 1.7(ii) below).

(iii) Continuous Functions: *For a given compact subset $\Omega \subset \mathbb{R}^m$ (see Definition 1.9(vi) below), let $C^0(\Omega)$ be the space of continuous functions $x : \Omega \to \mathbb{R}$ with supremum norm $\|\cdot\|_{C^0(\Omega)}$ defined by*

$$\|x\|_{C^0(\Omega)} = \sup_{u\in\Omega} |x(u)|,$$

for $x \in C^0(\Omega)$. Then, $(C^0(\Omega), \|\cdot\|_{C^0(\Omega)})$ is a normed vector space.

Example 1.5 (continued)**.**

(iv) L^p Spaces*: For some* $1 \leqslant p < \infty$ *and a measure space* (Ω, Σ, μ) *(e.g. a probability space such that* $\mu(\Omega) = 1$*), let* $L^p(\Omega) := L^p(\Omega, \Sigma, \mu)$ *be the space of functions* $x : \Omega \to \mathbb{R}$ *whose* $L^p(\Omega)$*-norm*

$$\|x\|_{L^p(\Omega)} = \left(\int_\Omega |x(u)|^p d\mu(u) \right)^{\frac{1}{p}}$$

is finite. Then, $(L^p(\Omega), \|\cdot\|_{L^p(\Omega)})$ *is a normed vector space and for* $p \leqslant q$*, the embedding* $L^q(\Omega) \subseteq L^p(\Omega)$ *holds true by the Hölder inequality (see Lemma 1.7(ii)). Moreover, case (ii) is a special case with* $\Omega = \mathbb{N}$ *and the counting measure* $\mu : \Sigma \to \mathbb{R}$ *defined by* $\mu(A) = \#A$*, for* $A \in \Sigma$*.*

Remark 1.6. *A significant difference between Example 1.5(ii)-(iv) and Example 1.5(i) is the dimension. While the Euclidean space* $\mathbb{R}^d$ *is finite dimensional with dimension* d*, the other spaces* l^p*,* $C^0(\Omega)$*, and* $L^p(\Omega)$ *are infinite dimensional (for* $C^0(\Omega)$ *and* $L^p(\Omega)$ *this is true if* Ω *or* $\operatorname{supp}(\mu)$ *are infinite dimensional sets). This can be seen by observing that a vector* $x \in \mathbb{R}^d$ *can be described by finitely many numbers, whereas a path* $x \in C^0([0, T])$ *contains infinite amount of information about the evolution of the trajectory.*

In the case of the Banach space $L^p(\Omega) := L^p(\Omega, \Sigma, \mu)$, the Minkowski inequality replaces the triangle inequality, where the Hölder inequality and Jensen inequality also play an important role.

Lemma 1.7. *For some* $1 \leqslant p < \infty$ *and a measure space* (Ω, Σ, μ)*, consider the Banach space* $L^p(\Omega) := L^p(\Omega, \Sigma, \mu)$*. Then, the following holds true.*

(i) Minkwoski's inequality: $\|f + g\|_{L^p(\Omega)} \leqslant \|f\|_{L^p(\Omega)} + \|g\|_{L^p(\Omega)}$*, for all* $f, g \in L^p(\Omega)$*.*

(ii) Hölder's inequality: $\|f \cdot g\|_{L^1(\Omega)} \leqslant \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}$*, for all* $f \in L^p(\Omega)$ *and* $g \in L^q(\Omega)$*, where* $1 < q \leqslant \infty$ *such that* $1/p + 1/q = 1$*.*

(iii) Jensen's inequality: If $\mu(\Omega) = 1$ *and if* $\varphi : \mathbb{R} \to \mathbb{R}$ *is convex, it holds that* $\varphi\left(\int_\Omega f(x) d\mu(x)\right) \leqslant \int_\Omega \varphi(f(x)) d\mu(x)$*, for all* $f \in L^1(\Omega)$*.*

In addition, a metric spaces can be either complete or incomplete. Let (X, d_X) be a metric and let $(x_n)_{n \in \mathbb{N}}$ be a sequence in X, which is called a *Cauchy sequence* if for every $\varepsilon > 0$ there exists some $N \in \mathbb{N}$ such that $d_X(x_n, x_m) < \varepsilon$,

for all $n, m \geqslant N$. On the other hand, the sequence $(x_n)_{n\in\mathbb{N}}$ is said to *converge* to the *limit* $x \in X$ if for every $\varepsilon > 0$ there exists some $N \in \mathbb{N}$ such that $d_X(x_n, x) < \varepsilon$, for all $n \geqslant N$.

Definition 1.8. *A metric space (X, d_X) is* complete *if every Cauchy sequence converges. Moreover, a complete normed vector space $(X, \|\cdot\|_X)$ is called a* Banach space*.*

We now introduce the topological properties of a given subset in a normed vector space $(X, \|\cdot\|_X)$.

Definition 1.9. *Let $(X, \|\cdot\|_X)$ be a normed vector space and define the* open ball of radius $r > 0$ with center $x_0 \in X$ *as $B_r(x_0) = \{x \in X;\, \|x - x_0\|_X < r\}$. Then, we introduce the following.*

(i) A subset $U \subset X$ is open *if for every $x_0 \in U$ there exists some $r > 0$ such that $B_r(x_0) \subset U$.*

(ii) A subset $A \subset X$ is closed *if $A^c := X\backslash A$ is open.*

(iii) The interior *of a subset $E \subset X$ is defined as $\mathring{E} = \bigcup_{G\subset E \text{ open}} G$, whereas its* closure *is defined as $\overline{E} = \bigcap_{E\subset G,\, G\subset X \text{ closed}} G$. In addition, the* boundary *of E is defined as $\partial E = \overline{E}\backslash\mathring{E}$.*

(iv) A subset $E \subset X$ is dense *if $\overline{E} = X$, i.e. if for every ball $B := B_r(x_0) \subset X$, it holds that $B \cap E \neq \{\}$.*

(v) A subset $E \subset X$ is bounded *if $\sup_{x\in E} \|x\|_X < \infty$.*

(vi) A subset $E \subset X$ is compact *if $E \subset X$ is closed and totally bounded, where the latter means that for every $\varepsilon > 0$ there exists finitely many points $x_1, ..., x_n \in X$ such that $X \subset \bigcup_{k=1}^{n} B_\varepsilon(x_k)$.*

These properties are all related to the norm topology, i.e. the topology induced by the norm $\|\cdot\|_X$.

Remark 1.10. *We give some examples of Definition 1.9. In $\mathbb{R}$ with metric $d_{\mathbb{R}}(x, y) = |x - y|$, the open interval $]a, b[$ with $a < b$ is an open subset, whereas $[a, b]$ is a closed subset. The interior of $E =]a, b]$ is equal to $]a, b[$, and the closure of E is given by $[a, b]$ such that the boundary is equal to $\{a, b\}$. In addition, it is known that the set of rational numbers $\mathbb{Q} = \left\{\pm\frac{p}{q};\, p, q \in \mathbb{N}\right\}$ is dense in $\mathbb{R}$. Finally, a subset of an Euclidean space is compact if and only if it is closed and bounded.*

If $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are Banach spaces, the space of continuous linear maps $T : X \to Y$ is denoted by $L(X;Y)$, where we note that not all linear maps are necessarily continuous in infinite dimensional spaces. However, the space $L(X;Y)$ defines again a Banach space with norm given by

$$\|T\|_{L(X;Y)} = \sup_{\substack{x\in X\\ x\neq 0}} \frac{\|Tx\|_Y}{\|x\|_X} = \sup_{\substack{x\in X\\ \|x\|_X\leqslant 1}} \|Tx\|_Y,$$

for $T \in L(X;Y)$. In the real-valued case of $Y = \mathbb{R}$, the space $X^* := L(X;\mathbb{R})$ is called the *dual space of* X consisting of continuous linear functionals on X. Hereby, the following Hahn-Banach theorem plays a crucial role together with the well-known separation theorems in [Str14, Satz 4.2.4+4.5.1], which can be used to show the following useful result in Corollary 1.14 for the machine learning applications of this thesis.

Theorem 1.11 (Hahn-Banach, [Str14, Satz 4.1.1])**.** *Let X be a vector space with a vector subspace $E \subset X$ and suppose that $p : X \to \mathbb{R}$ is a sublinear function, i.e.*

$$\begin{aligned} p(\lambda x) = \lambda p(x), \quad &\text{for all } \lambda \geqslant 0, \qquad \text{and} \\ p(x+y) \leqslant p(x) + p(y), \quad &\text{for all } x, y \in X. \end{aligned} \tag{1.2.2}$$

If $f : E \to \mathbb{R}$ is a linear map such that $f(x) \leqslant p(x)$, for all $x \in E$, then there exists a linear extension $F : X \to \mathbb{R}$ such that $F|_E = f$ and $F(x) \leqslant p(x)$, for all $x \in X$.

Remark 1.12. *Every norm satisfies the two conditions in* (1.2.2) *and is therefore a sublinear function.*

Theorem 1.13 (Hahn-Banach Separation, [Str14, Satz 4.2.4])**.** *Let $E \subsetneq X$ be a proper closed vector subspace of a Banach space $(X, \|\cdot\|_X)$ and assume that $x_0 X \backslash E$ with $d := \operatorname{dist}(x_0, E) = \inf_{x\in E} \|x - x_0\|_X > 0$. Then, there exists some $l \in X^*$ such that $l(x) = 0$, for all $x \in E$, and $l(x_0) = d$.*

Proof. Let $E_0 = \{x + tx_0;\, x \in E,\, t \in \mathbb{R}\}$ and define the linear function $f : E_0 \to \mathbb{R}$ by $f(x + tx_0) = td$ such that $f(x) = 0$, for all $x \in E$, and $f(x_0) = d > 0$. For $y = x + tx_0 \in E_0$ with $t \neq 0$, it follows that

$$|f(y)| = |t|d \leqslant |t| \left\| x_0 - \frac{-x}{t} \right\|_X = \|tx_0 + x\|_X = \|y\|_X,$$

such that $\|f\|_{L(E_0;\mathbb{R})} \leqslant 1$. In order to show that $\|f\|_{L(E_0;\mathbb{R})} = 1$, choose some $\varepsilon > 0$ and $x \in E$ with $d \leqslant \|x_0 - x\|_X < d + \varepsilon$. Then, it follows for $y = x_0 - x \in E_0$ that $f(y) = d \geqslant \frac{d}{d+\varepsilon}\|y\|_X$, which shows that

$$\|f\|_{L(E_0;\mathbb{R})} = \sup_{0 \neq y \in E_0} \frac{|f(y)|}{\|y\|_X} \geqslant \frac{d}{d+\varepsilon}.$$

Letting $\varepsilon \to 0$ implies that $\|f\|_{L(E_0;\mathbb{R})} = 1$. By applying Theorem 1.11, we obtain the linear functional $l \in X^*$ with $l(x) = 0$, for all $x \in E$, and $l(x_0) = d$ as extension of $f : E_0 \to \mathbb{R}$ to X. □

Corollary 1.14 (Hahn-Banach Denseness)**.** *Let $E \subset X$ be a vector subspace of a Banach space $(X, \|\cdot\|_X)$. Then, E is dense in X if and only if for every continuous linear functional $l : X \to \mathbb{R}$ vanishing on E, i.e. $l(x) = 0$, for all $x \in E$, it follows that $l : X \to \mathbb{R}$ vanishes everywhere, i.e. $l(x) = 0$, for all $x \in X$.*

Proof. We prove that $\overline{E} \subset X$ is a proper closed vector subspace if and only if there exists a continuous linear functional $l \in X^*$ vanishing on E, which is not identical to zero everywhere on X.

For necessity, let $l \in X^*$ be a non-zero continuous linear functional vanishing on E. By continuity it follows that $l : X \to \mathbb{R}$ vanishes on $\overline{E}$, which is a proper subset of X, because $l : X \to \mathbb{R}$ is non-zero.

For sufficiency, let $\overline{E}$ be a proper closed vector subspace of X such that there exists a point $x_0 \in X \backslash E$ with $\operatorname{dist}(x_0, E) := \inf\{\|x - x_0\|_X;\, x \in E\} > 0$. By the Hahn-Banach separation theorem in Theorem 1.13, we obtain a continuous linear functional $l \in X^*$ with $l(x_0) = \operatorname{dist}(x_0, E) > 0$, $l(x) = 0$, for all $x \in E$, and $\|l\|_{X*} = 1$. Hence, $l : X \to \mathbb{R}$ vanishes on E, but is not identical to zero everywhere on X. □

Using the dual space of a normed vector space $(X, \|\cdot\|_X)$, we can define weaker topologies than the norm topology on $(X, \|\cdot\|_X)$ and $(X^*, \|\cdot\|_{X*})$. On $(X, \|\cdot\|_X)$, the *weak topology* τ_w is induced by weak convergence, where a sequence $(x_n)_{n\in\mathbb{N}}$ in X *converges weakly* to some $x \in X$ if $l(x_n) \to l(x)$, for all $l \in X^*$, as $n \to \infty$. On the other hand, the *weak-$*$-topology* τ_{w*} on X^* is induced by the weak-$*$-convergence, where a sequence $(l_n)_{n\in\mathbb{N}}$ in X^* *converges in the weak-$*$-topology* to some $l \in X^*$ if $l_n(x) \to l(x)$, for all $x \in X$, as $n \to \infty$. Both topologies are in general not metrizable unless X is finite dimensional.

In the following, we now give different theorems to verify compactness of a set in a given topology, which is in particular useful for machine learning applications. We start with the well-known Banach-Alaoglu theorem involving the weak-$*$-topology, followed by Arzelà-Ascoli results for the norm topology.

Theorem 1.15 (Banach-Alaoglu, [Rud91, Theorem 3.15]). *Let X be the dual of a normed vector space $(S, \|\cdot\|_S)$. Then, the unit ball $\{x \in X;\ \|x\|_{S^*} \leqslant 1\}$ is compact with respect to the weak-$*$-topology of X.*

For the rest of this section, we assume that (X, d_X) is a compact metric space and $(Y, \|\cdot\|_Y)$ is a Banach space such that $C^0(X;Y)$ denotes the vector space of Y-valued continuous functions with supremum norm $\|f\|_{C^0(X;Y)} = \sup_{x\in X} \|f(x)\|_Y$, where $E \subset C^0(X;Y)$ is called *relatively compact* if $\overline{E}$ is compact in $C^0(X;Y)$.

Theorem 1.16 (Arzelà-Ascoli, [Rud91, Appendix A5]). *Let (X, d_X) be a compact metric space and let $(Y, \|\cdot\|_Y)$ be a Banach space. Then, $E \subset C^0(X;Y)$ is relatively compact if and only if E is equicontinuous, i.e.*

$$\forall \varepsilon > 0 : \exists \delta > 0 : \forall f \in E : \forall x, y \in X : \quad d_X(x,y) < \delta \Rightarrow \|f(x) - f(y)\|_Y < \varepsilon,$$

and there exists a dense subset $D \subset X$ such that $\{f(x);\ f \in E\}$ is relatively compact in Y, for all $x \in D$.

Corollary 1.17 (Arzelà-Ascoli, [Str14, Satz 6.3.1]). *Let $\Omega \subset \mathbb{R}^m$ be a bounded subset. Then, $E \subset C^0(\overline{\Omega})$ is relatively compact if and only if E is bounded and equicontinuous, i.e. $\sup_{f\in E} \|f\|_{C^0(\Omega)} < \infty$ and*

$$\forall \varepsilon > 0 : \exists \delta > 0 : \forall f \in E : \forall x, y \in \Omega : \quad \|x - y\| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon.$$

Corollary 1.18 (Fréchet-Kolmogorov, [Str14, Satz 6.3.2]). *Let $1 \leqslant p < \infty$ and $\Omega \subset \mathbb{R}^m$ be bounded, and denote by $\lambda : \mathcal{F} \to \mathbb{R}$ the Lebesgue measure, where $\mathcal{F}$ denotes the Borel σ-algebra of $\mathbb{R}^m$. Then, $E \subset L^p(\Omega) := L^p(\Omega, \mathcal{F}, \lambda)$ is relatively compact if and only if E is bounded and equicontinuous in L^p, i.e. $\sup_{f\in E} \|f\|_{L^p(\Omega)} < \infty$ and*

$$\sup_{f\in E} \|f - f(\cdot + h)\|_{L^p(\mathbb{R}^m)} \quad \overset{h\to 0}{\longrightarrow} \quad 0,$$

where $f : \Omega \to \mathbb{R}$ is extended to all of $\mathbb{R}^m$ by setting it to zero outside of Ω.

These results cover the most preliminary mathematical results for metric spaces and Banach spaces, which can be extended in some cases to more general topological space introduced in the next section.

1.2.2 Topological Spaces and Convenient Vector Spaces

In order to consider more general spaces than metric spaces and Banach spaces, we introduce topological spaces in the following. They are considered to be the most general type of mathematical spaces, which allow for the definitions of limits, continuity and connectedness, see [Kel75].

Definition 1.19. *Let X be a non-empty set and let τ_X be a collection of subsets of X. Then, τ_X defines a* topology *on X if the following three axioms are satisfied:*

(i) Empty and full set: $\{\}, X \in \tau_X$,

(ii) Arbitrary unions: *for all $(U_i)_{i\in I}$ with arbitrary I and $U_i \in \tau_X$, it follows that $\bigcup_{i\in I} U_i \in \tau_X$,*

(iii) Finite intersections: *for all $(U_i)_{i=1}^n$ with some $n \in \mathbb{N}_{\geqslant 1}$ and $U_i \in \tau_X$, it follows that $\bigcap_{i=1}^n U_i \in \tau_X$.*

If τ_X is a topology on X, (X, τ_X) is called a topological space*, where τ_X describes the open sets. Furthermore, $E \in \tau_X$ is an* open neighborhood *of $x \in X$ if $x \in E$. A* neighborhood basis of $x \in X$ *is a subfamily $\tau_x \subset \tau$ of open neighborhoods of x such that for any open neighborhood E of x there is some $U \in \tau_x$ with $U \subset E$.*

Example 1.20. *On the set $X = \{1, 2\}$, the collection $\tau_X = \{\{\}, \{1, 2\}\}$ defines the so-called* trivial topology*, where only the empty and full set are open. On the other hand, taking the power set as collection $\tau_X = 2^X := \{\{\}, \{1\}, \{2\}, \{1, 2\}\}$ defines the so-called* discrete topology*, where every subset is open and closed.*

Analogously to Definition 1.9, we repeat the topological properties of subsets in a topological space.

Definition 1.21. *Let (X, τ_X) be a topological space. Then, we introduce the following.*

(i) A subset $U \subset X$ is called open *if $U \in \tau_X$.*

(ii) A subset $A \subset X$ is called closed *if $A^c := X \backslash A \in \tau_X$.*

(iii) The interior *of a subset $E \subset X$ is defined as $\mathring{E} = \bigcup_{G\subset E,\, G\in\tau_X} G$, whereas its* closure *is defined as $\overline{E} = \bigcap_{E\subset G,\, G^c\in\tau_X} G$. In addition, the* boundary *of E is defined as $\partial E = \overline{E} \backslash \mathring{E}$.*

Definition 1.21 (continued)**.**

(iv) A subset $E \subset X$ is called dense *if $\overline{E} = X$.*

(v) A subset $E \subset X$ is called compact *if every open cover of E has a finite subcover, i.e. if for every $(U_i)_{i\in I}$ with $U_i \in \tau_X$ and $E \subset \bigcup_{i\in I} U_i$, we find some finite index set $I_0 \subset I$ such that $E \subset \bigcup_{i\in I_0} U_i$.*

These properties are all related to the given topology τ_X. Moreover, if (X, τ_1) and (X, τ_2) are topological spaces with $\tau_1 \subseteq \tau_2$, then τ_1 (resp. τ_2) is called a weaker (resp. stronger) topology *than τ_2 (resp. τ_1).*

Moreover, a function $f : X \to Y$ between two topological spaces (X, τ_X) and (Y, τ_Y) is called *continuous* if for every open subset $V \in \tau_Y$, the preimage $f^{-1}(V) := \{x \in X;\, f(x) \in V\} \in \tau_X$ is open as well. Using this, we introduce the following separation axioms for topological spaces.

Definition 1.22. *Let (X, τ_X) be a topological space. Then, we introduce the following:*

(i) (X, τ_X) is Hausdorff *if for every pair of distinct points $x, y \in X$ there exist open neighborhoods U and V of x and y, respectively, such that $U \cap V = \{\}$.*

(ii) (X, τ_X) is completely regular *if for every point $x_0 \in X$ and every closed subset $A \subset X$ with $x_0 \notin A$ there exists a continuous function $f : X \to \mathbb{R}$ such that $f(x_0) = 1$ and $f(x) = 0$, for all $x \in A$.*

A completely regular Hausdorff space is also called a Tychonoff space*.*

A special class of topological spaces is given by topological vector spaces, which satisfy the vector space axioms and for which vector addition $X \times X \ni (x, y) \mapsto x + y \in X$ and scalar multiplication $\mathbb{R} \times X \ni (\lambda, x) \mapsto \lambda \cdot x \in X$ are continuous. A subclass is formed by the topological vector spaces, whose topology is locally convex. Hereby, a subset $E \subset X$ is called *convex* if $\lambda x + (1 - \lambda)y \in E$, for all $x, y \in X$ and $\lambda \in [0, 1]$.

Definition 1.23. *A Hausdorff topological vector space (X, τ_X) is locally convex if the origin has a neighborhood basis, usually denoted by τ_0, which consists solely of convex sets.*

In the following, we shall see that the topology of locally convex topological vector spaces can be characterised by seminorms, where we assume from now on that (X, τ_X) denotes a topological vector space.

Definition 1.24. *The map $p : X \to \mathbb{R}$ is called a* seminorm *if the following properties are satisfied:*

(i) Absolute homogeneity: $p(\lambda x) = |\lambda| p(x)$, *for all* $\lambda \in \mathbb{R}$ *and* $x \in X$,

(ii) Subadditivity: $p(x + y) \leqslant p(x) + p(y)$, *for all* $x, y \in X$.

Moreover, if a seminorm $p : X \to \mathbb{R}$ is positive definite, i.e. $p(x) = 0 \Leftrightarrow x = 0$, then $p : X \to \mathbb{R}$ is a norm.

In addition, a subset $E \subset X$ is called *balanced* if $\lambda E := \{\lambda x;\, x \in E\} \subseteq E$, for all $\lambda \in [-1, 1]$, whereas a subset $E \subset X$ is *absorbing* if for every $x \in X$ there exists some $t > 0$ such that $|\lambda| \geqslant t$ if and only if $x \in \lambda E$.

Theorem 1.25 ([SW99, Chapter II§1])**.** *If $\mathcal{P}$ is a family of seminorms on X, then the topology τ_X induced by sets of the form $\{x \in X;\, p(x) < \varepsilon\}$, for $\varepsilon > 0$ and $p \in \mathcal{P}$, turns (X, τ_X) into a locally convex topological vector space. Conversely, if (X, τ_X) is a locally convex topological vector space with neighborhood base of zero τ_0 consisting of balanced and convex sets, then the locally convex topology induced by the Minkowski functionals $p_E(x) = \inf\{\lambda > 0;\, x \in \lambda E\}$, for $E \in \tau_0$, coincides with τ_X.*

Example 1.26. *For an open subset $\Omega \subset \mathbb{R}^n$, let $X = C_c^\infty(\Omega)$ be the space of compactly supported smooth functions $x : \Omega \to \mathbb{R}$ equipped with the topology τ_X, which is generated by the seminorms $p_{k,K}(x) = \max_{|\alpha| \leqslant k} \sup_{u \in K} |\partial^\alpha x(u)|$, for $k \in \mathbb{N}$ and compact subsets $K \subset \Omega$, where $\alpha = (\alpha_1, ..., \alpha_n) \in \mathbb{N}^n$ denotes a multi-index with $|\alpha| := \alpha_1 + ... + \alpha_n$. Then, $(C_c^\infty(\Omega), \tau_X)$ defines a locally convex topological vector space.*

For the definition of completeness of a Hausdorff topological vector space (X, τ_X), we define a *net* as a pair (A, f), where $f : A \to X$ is a map from the directed set $(A, \leqslant)$ to X, which converges to some point $x \in X$ if for every open neighborhood E of x there exists some $a \in A$ such that $f(b) \in E$, for all $b \geqslant a$. In addition, a net (A, f) is a *Cauchy net* if for every neighborhood E of zero there exists some $a \in A$ such that $f(b) - f(c) \in E$, for all $b, c \geqslant a$.

Definition 1.27. *A Hausdorff topological vector space (X, τ_X) is said to be* complete *if every Cauchy net (A, f) converges to some point $x \in X$.*

The completion of a Hausdorff topological vector space (X, τ_X) is given by the complete topological vector space (Y, τ_Y) such that $\iota : X \to Y$ defines a continuous linear embedding with dense image and τ_X coincides with the topology

$\{\iota^{-1}(V);\, V \in \tau_Y\}$. Moreover, one can show that every Hausdorff topological vector space possesses a completion, which is unique up to isomorphism, see [SW99, Theorem I§1.5].

Now, we consider the dual space X^* consisting of all continuous linear functionals $l : X \to \mathbb{R}$, on which we extend the notion of the weak and weak-$*$-topology. The *weak topology* is the weakest topology on X such that all maps $X \ni x \mapsto l(x) \in \mathbb{R}$ are continuous, for all $l \in X^*$, whereas the *weak-$*$-topology* is the weakest topology on X^* such that all maps $X^* \ni l \mapsto l(x) \in \mathbb{R}$ are continuous, for all $x \in X$.

Theorem 1.28 (Hahn-Banach, [SW99, Theorem II§4.2])**.** *Let (X, τ_X) be a locally convex topological vector space and let $E \subset X$ be a vector subspace. Suppose that $f : E \to \mathbb{R}$ is a continuous linear functional on E. Then, there exists a continuous linear functional $F : X \to \mathbb{R}$ such that $F|_E = f$.*

Corollary 1.29. *Let (X, τ_X) be a locally convex topological vector space and let $E \subset X$ be a vector subspace. Then, E is dense in X if and only if for every continuous linear functional $l : X \to \mathbb{R}$ vanishing on E, it follows that $l : X \to \mathbb{R}$ vanishes everywhere.*

Proof. The proof follows along the lines of the proof of Corollary 1.14, where we use the Hahn-Banach theorem for locally convex topological vector spaces in Theorem 1.28 instead of Theorem 1.11. □

In order to define differentiability on locally convex topological vector spaces, we use the satisfactory and natural approach of convenient analysis introduced in [KM97], where we assume that (X, τ_X) denotes a locally convex topological vector space, whose topology τ_X is generated by the family of seminorms $\mathcal{P}$. Hereby, a subset $E \subset X$ is *bounded* if $p(E) := \{p(x);\, x \in E\}$ is bounded in $\mathbb{R}$, for all seminorms $p \in \mathcal{P}$.

A sequence $(x_n)_{n\in\mathbb{N}} \subset X$ is called *Mackey-convergent* to the limit $x \in X$ if there exists a bounded, convex and balanced subset $E \subset X$ and a sequence $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ with $\lim_{n\to\infty} a_n = 0$ such that $x_n - x \in a_n E$, whereas $(x_n)_{n\in\mathbb{N}}$ is called a *Mackey-Cauchy sequence* if there exists a bounded, convex and balanced subset $E \subset X$ and a double sequence $(a_{nm})_{n,m\in\mathbb{N}}$ with $\lim_{n,m\to\infty} a_{nm} = 0$ such that $x_n - x_m \in a_{nm} E$.

Definition 1.30. *A locally convex topological vector space (X, τ_X) is called* convenient *if every Mackey-Cauchy sequence $(x_n)_{n\in\mathbb{N}} \subset X$ converges to some limit $x \in X$.*

For a convenient vector space (X, τ_X), the vector space of smooth curves $C^\infty(\mathbb{R}; X)$ does not depend entirely on the topology τ_X, only on the corresponding bornology (system of bounded sets), see [KM97, Corollary I.2.11]. Hereby, a curve $c : \mathbb{R} \to X$ is called *differentiable* if the limit $c'(t) := \lim_{h\to 0} \frac{c(t+h)-c(t)}{h}$ exists, for all $t \in \mathbb{R}$, whereas a curve $c : \mathbb{R} \to X$ is *smooth* if the iterated derivatives $c^{(k+1)} = \left(c^{(k)}\right)'$ exist, for all $k \in \mathbb{N}$. In addition, we call $c : \mathbb{R} \to X$ a $\mathcal{L}ip^k$*-curve* if all derivatives up to order k exist and are locally Lipschitz, i.e. every point $r \in \mathbb{R}$ has a neighborhood $U \subset \mathbb{R}$ such that $\left\{ \frac{c(t)-c(s)}{t-s};\, t, s \in U,\, t \neq s \right\}$ is bounded. Then, convenient vector spaces are equivalently characterized by the differentiability of X-valued curves.

Theorem 1.31 ([KM97, Theorem I.2.14])**.** *The following are equivalent:*

(i) (X, τ_X) *is convenient.*

(ii) For every curve $c : \mathbb{R} \to X$ *such that* $l \circ c$ *is smooth, for all* $l \in X^*$*, it follows that* $c \in C^\infty(\mathbb{R}, X)$*.*

(iii) Any smooth curve $c \in C^\infty(\mathbb{R}; X)$ *is locally Riemann-integrable.*

By [KM97, Lemma I.2.2+Proposition I.3.7] and assertion (iii), every sequentially complete locally convex vector spaces is convenient, whereas condition (ii) in Theorem 1.31 ensures that the concept of weakly smooth curves and smooth curves are equivalent in convenient vector spaces. Furthermore, we define the c^∞*-topology* τ_{c^∞} on a convenient vector space (X, τ_X) as the final topology with respect to all smooth curves $c \in C^\infty(\mathbb{R}; X)$ such that $E \subset X$ is c^∞*-open* if and only if $c^{-1}(E)$ is open in $\mathbb{R}$, for all $c \in C^\infty(\mathbb{R}; X)$. Then, it follows by [KM97, Theorem I.2.14] that the c^∞-topology is finer than any locally convex topology on X with the same bounded subsets such that the embedding $(X, \tau_{c^\infty}) \subseteq (X, \tau_X)$ holds true, where (X, τ_{c^∞}) is not necessarily a topological vector space because the continuity of the vector addition might be violated.

For convenient vector spaces (X, τ_X), (Y, τ_Y) and a c^∞-open subset $E \subset X$, a map $f : E \to Y$ is called *smooth (on* E*)* if for every smooth curve $c \in C^\infty(\mathbb{R}; E)$, the curve $f \circ c \in C^\infty(\mathbb{R}; Y)$ is also smooth. We denote by $C^\infty(E; Y)$ the vector spaces of smooth maps $f : E \to Y$ equipped with the initial locally convex topology of $C^\infty(E; Y) \ni f \mapsto l \circ f \circ c \in C^\infty(\mathbb{R}; \mathbb{R})$, for $c \in C^\infty(\mathbb{R}; E)$ and $Y' := \{l : Y \to \mathbb{R} \text{ linear and bounded}\}$. This topology is the weakest topology on $C^\infty(E; Y)$ such that $f \mapsto l \circ f \circ c$ is continuous, for all $c \in C^\infty(\mathbb{R}; E)$ and $l \in Y'$, where $C^\infty(\mathbb{R}; \mathbb{R})$ is equipped with the topology of compact convergence in each derivative. Then, $C^\infty(E; Y)$ is by [KM97, Lemma I.3.7+I.3.11] a convenient vector space and we define $C^\infty(E) := C^\infty(E; \mathbb{R})$.

Similarly, a map $f : E \to Y$ is called *$\mathcal{L}ip^k$ (on E)* if $f \circ c \in \mathcal{L}\mathrm{ip}^k(\mathbb{R}; Y)$, for all $c \in C^\infty(\mathbb{R}; E)$, where $\mathcal{L}\mathrm{ip}^k(\mathbb{R}; Y)$ is the vector space of $\mathcal{L}\mathrm{ip}^k$-curves equipped with the initial topology of the derivatives of order $j \leqslant k$ considered as linear mappings into the space of $\mathcal{L}\mathrm{ip}^0$-curves equipped with the locally convex topology of uniform convergence on bounded subsets of $\mathbb{R}$ and of the difference quotient on bounded subsets of $\big\{(s,t) \in \mathbb{R}^2;\, s \neq t\big\}$, see [KM97, Definition III.12.10]. Using this, we denote by $\mathcal{L}\mathrm{ip}^k(E; Y)$ the vector space of $\mathcal{L}\mathrm{ip}^k$-maps $f : E \to Y$ with the initial topology of $\mathcal{L}\mathrm{ip}^k(E; Y) \ni f \mapsto f \circ c \in \mathcal{L}\mathrm{ip}^k(\mathbb{R}; Y)$, for $c \in C^\infty(\mathbb{R}; E)$. By [KM97, Definition III.12.11], $\mathcal{L}\mathrm{ip}^k(E; Y)$ is a convenient vector space and we define $\mathcal{L}\mathrm{ip}^k(E) := \mathcal{L}\mathrm{ip}^k(E; \mathbb{R})$.

Theorem 1.32 (Main Structure Theorem for Convenient Spaces, [KM97] and [FK88])**.** *Let (X, τ_X), (Y, τ_Y), (Z, τ_Z) be convenient vector spaces and let $E \subset X$ and $F \subset Y$ be c^∞-open. Then, the following hold true.*

(i) Every smooth map $f \in C^\infty(E; Y)$ is c^∞-continuous.

(ii) Multilinear maps $f : X \times ... \times X \to Y$ are smooth if and only if they are bounded.

(iii) For a smooth map $f \in C^\infty(E; Y)$ (resp. $\mathcal{L}ip^{k+1}$-map $f \in \mathcal{L}ip^{k+1}(E; Y)$), the directional derivative $df : E \times X \to Y$ defined by

$$df(x)(v) = \frac{d}{dt}\Big|_{t=0} f(x + tv),$$

for $x \in E$ and $v \in X$, exists, is linear, bounded and smooth (resp. $\mathcal{L}ip^k$). Similarly, the higher order derivatives $d^2 f : E \times X \times X \to Y$, ..., $d^k f : E \times X \times ... \times X \to Y$ exist, which is true for any $k \in \mathbb{N}$ in the case of $f \in C^\infty(E; Y)$.

(iv) The chain rule holds true, i.e. for smooth maps $g \in C^\infty(E; Y)$ and $f \in C^\infty(F; Z)$ (resp. $\mathcal{L}ip^k$-maps $g \in \mathcal{L}ip^k(E; Y)$ and $f \in \mathcal{L}ip^k(F; Z)$, where $k \geqslant 1$) with $g(E) \subset F$, it holds that $d(f \circ g)(x)(v) = df(g(x))(dg(x)(v))$, for all $x \in E$ and $v \in X$.

(v) Taylor's formula holds true, i.e. for every smooth map $f \in C^\infty(E; Y)$ (resp. $f \in \mathcal{L}ip^{k+1}(E; Y)$), it holds that

$$f(x + v) = \sum_{j=0}^{n} \frac{1}{j!} d^j f(x)(v, ..., v) + \int_0^1 \frac{(1-t)^n}{n!} d^{n+1} f(x + tv)(v, ..., v) dt,$$

for all $n \in \mathbb{N}$ (resp. $n \leqslant k$) and $[x, x + v] := \{x + tv;\, t \in [0, 1]\} \subset E$.

Theorem 1.32 (continued)**.**

(vi) The exponential law holds true (cartesian closedness), i.e. $C^\infty(E\times F;Z)\cong C^\infty(E;C^\infty(F;Z))$.

(vii) The evaluation map $\mathrm{ev}:C^\infty(E;Y)\times X\to Y$ *defined by* $\mathrm{ev}(f,x)=f(x)$ *and the composition map* $\mathrm{comp}:C^\infty(Y;Z)\times C^\infty(E;Y)\to C^\infty(E;Z)$ *defined by* $\mathrm{comp}(f,g)=f\circ g$ *are smooth.*

(viii) A linear map $f:E\to C^\infty(F;Z)$ *is smooth if and only if* $\mathrm{ev}(\cdot,v)\circ f:F\to Z$ *is smooth, for all* $v\in F$.

Theorem 1.32 shows that convenient calculus inherits many analytical properties, which are true on finite dimensional spaces, making the theory natural for applications in infinite dimensional vector spaces. Further details on convenient calculus can be found in [KM97] and the references therein.

1.2.3 Stone-Weierstrass and Nachbin's Theorem

The Stone-Weierstrass theorem and Nachbin's result can be used to establish density of a function class in a given function space. We begin with the vector space of polynomials over a subset $E\subset\mathbb{R}$ collected in

$$\mathrm{Pol}(E)=\left\{E\ni x\mapsto\sum_{k=0}^{n}a_kx^k;\ n\in\mathbb{N},\ a_k\in\mathbb{R}\right\}.$$

For a compact subset $K\subset\mathbb{R}$, Karl Weierstrass proved in [Wei85] that polynomials in $\mathrm{Pol}(K)$ are dense in $C^0(K)$, which we show in the following with the simple and beautiful proof of Sergei Bernstein. For this purpose, we use the Bernstein polynomials defined by $B_{n,k}(x)=\binom{n}{k}x^k(1-x)^{n-k}$, for $x\in\mathbb{R}$, $n\in\mathbb{N}$ and $k=0,...,n$ and introduce the Bernstein approximation operator $B_n:C^0([0,1])\to C^0([0,1])$ defined by

$$(B_nf)(x)=\sum_{k=0}^{n}f\left(\frac{k}{n}\right)B_{n,k}(x)=\sum_{k=0}^{n}f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k},$$

for $f\in C^0([0,1])$ and $x\in[0,1]$, where $B_n(f)\in\mathrm{Pol}([0,1])$ is a polynomial approximating $f\in C^0([0,1])$.

Theorem 1.33 (Weierstrass, [Wei85])**.** *Let* $K\subset\mathbb{R}$ *be compact. Then,* $\mathrm{Pol}(K)$ *is dense in* $C^0(K)$.

Proof. By using affine transformations, we can restrict ourselves without loss of generality to $K = [0,1]$. Let $x \in K$ and assume that $(X_k)_{k\in\mathbb{N}_{\geqslant 1}}$ is an i.i.d. sequence of Bernoulli-distributed random variable such that $\mathbb{P}[X_k = 0] = 1-x$ and $\mathbb{P}[X_k = 1] = x$. Define $S_n := X_1 + ... + X_n$, for which the law of large numbers shows that $\frac{1}{n}S_n$ converges in probability to $\mathbb{E}[X_k] = x$, as $n \to \infty$, where S_n is binomially distributed, i.e.

$$\mathbb{P}[S_n = k] = \binom{n}{k} x^k (1-x)^{n-k} = B_{n,k}(x),$$

for $k = 0, ..., n$. Hence it holds for some given $f \in C^0([0,1])$ and $\delta > 0$ that

$$\begin{aligned} |(B_n f)(x) - f(x)| &\leqslant \mathbb{E}\left[\left| f\left(\frac{S_n}{n}\right) - f(x)\right|\right] \leqslant 2\mathbb{P}\left[\left|\frac{S_n}{n} - x\right| \geqslant \delta\right] \sup_{u\in[0,1]} |f(u)| \\ &\quad + \mathbb{P}\left[\left|\frac{S_n}{n} - x\right| < \delta\right] \sup_{|u-v|<\delta} |f(u) - f(v)|. \end{aligned} \tag{1.2.3}$$

Since $f \in C^0([0,1])$ is uniformly continuous, there exists for fixed $\varepsilon > 0$ some $\delta > 0$ such that $|u - v| < \delta$ implies $|f(u) - f(v)| < \varepsilon/2$, for all $u, v \in [0,1]$. Using this for the second term on the right-hand side of (1.2.3), we have

$$\mathbb{P}\left[\left|\frac{S_n}{n} - x\right| < \delta\right] \sup_{|u-v|<\delta} |f(u) - f(v)| \leqslant \sup_{|u-v|<\delta} |f(u) - f(v)| < \frac{\varepsilon}{2}. \tag{1.2.4}$$

On the other hand, using Markov's inequality for the first term on the right-hand side of (1.2.3), inserting $\mathbb{E}\left[(S_n - nx)^2\right] = nx(1-x) \leqslant n/4$ and choosing $n \in \mathbb{N}_{\geqslant 1}$ such that $n > \|f\|_{C^0([0,1])}/(\delta^2\varepsilon)$, it follows that

$$\begin{aligned} 2\mathbb{P}\left[\left|\frac{S_n}{n} - x\right| \geqslant \delta\right] \sup_{u\in[0,1]} |f(u)| &\leqslant \frac{2}{\delta^2}\mathbb{E}\left[\left(\frac{S_n}{n} - x\right)^2\right] \|f\|_{C^0([0,1])} \\ &\leqslant \frac{1}{2n\delta^2} \sup_{u\in[0,1]} \|f\|_{C^0([0,1])} < \frac{\varepsilon}{2}. \end{aligned} \tag{1.2.5}$$

By inserting (1.2.4) and (1.2.5) into (1.2.3), we conclude for large enough $n \in \mathbb{N}$ that

$$\|B_n f - f\|_{C^0([0,1])} = \sup_{x\in[0,1]} |(B_n f)(x) - f(x)| < \varepsilon.$$

Since $\varepsilon > 0$ and $f \in C^0([0,1])$ were chosen arbitrarily, it follows that $\mathrm{Pol}([0,1])$ is dense in $C^0([0,1])$. □

Subsequently, Marshall Harvey Stone generalized Weierstrass' approximation result in [Sto48] to more general compact Hausdorff topological spaces by replacing the polynomials with an arbitrary subalgebra.

Definition 1.34. *Let (X, τ_X) be a topological Hausdorff space and let $\mathcal{A} \subset C^0(X)$ be a subset. Then, we introduce the following.*

(i) $\mathcal{A}$ *is* closed under addition (or additively closed) *if $a_1 + a_2 \in \mathcal{A}$, for all $a_1, a_2 \in \mathcal{A}$.*

(ii) $\mathcal{A}$ *is* closed under multiplication (or multiplicatively closed) *if $a_1 \cdot a_2 \in \mathcal{A}$, for all $a_1, a_2 \in \mathcal{A}$, where the multiplication is defined pointwise by $(a_1 \cdot a_2)(x) = a_1(x)\, a_2(x)$, for every $x \in X$.*

(iii) $\mathcal{A}$ *is* point separating (on X) *if for distinct $x_1, x_2 \in X$ there exists some $a \in \mathcal{A}$ with $a(x_1) \neq a(x_2)$.*

(iv) $\mathcal{A}$ vanishes nowhere (on X) *if for every $x \in X$ there exists some $a \in \mathcal{A}$ with $a(x) \neq 0$.*

Moreover, a vector subspace $\mathcal{A} \subset C^0(X)$ is called a subalgebra (of $C^0(X)$) *if $\mathcal{A}$ is closed under multiplication.*

Theorem 1.35 (Stone-Weierstrass, [Sto48])**.** *Let (K, τ_K) be a compact Hausdorff topological space and assume that $\mathcal{A} \subset C^0(K)$ is a subalgebra. Then, $\mathcal{A}$ is dense in $C^0(K)$ if and only if $\mathcal{A}$ is point separating and vanishes nowhere.*

Indeed, Theorem 1.35 generalizes Theorem 1.33 since $\mathrm{Pol}(K)$ is a point separating subalgebra of $C^0(K)$, where the sum and product of two polynomials define again a polynomial and the identity $x \mapsto p(x) := x \in \mathrm{Pol}(K)$ separates points of K because we have $p(x_1) = x_1 \neq x_2 = p(x_2)$, for distinct points $x_1, x_2 \in K$. Later, the Stone-Weierstrass theorem has been generalized to the vector-valued case, where the output is infinite dimensional (see [Pro77]), and to non-compact input spaces using weighted topologies (see [Nac65]).

Leopoldo Nachbin extended the Stone-Weierstrass theorem in [Nac49] to the approximation of the derivatives. For $E \subset \mathbb{R}^d$ open, a function $f : E \to \mathbb{R}$ is *k-times differentiable (on E)* if the partial derivatives $\partial^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ exist on E and are continuous, for all multi-indices $\alpha \in \mathbb{N}_k^d := \big\{\alpha \in \mathbb{N}^d;\, |\alpha| := \alpha_1 + \ldots + \alpha_d \leqslant k\big\}$. From this, we define $C^k(E)$ as the vector space of k-times differentiabl functions on E equipped with compact-open

topology of order k, which is generated by seminorms of the form

$$\|f\|_{C^k(K)} = \sum_{\alpha \in \mathbb{N}_k^d} \|\partial^\alpha f\|_{C^0(K)} = \sum_{\alpha \in \mathbb{N}_k^d} \sup_{x \in K} |\partial^\alpha f(x)|,$$

for $K \subset E$ compact. Hereby, the Bernstein polynomials $B_{n,k}$ can also be used to approximate a function $f \in C^k(E)$ including its derivatives, which generalizes Theorem 1.33.

Theorem 1.36 (Weierstrass with Derivatives)**.** *Let $E \subset \mathbb{R}$ be open. Then,* $\mathrm{Pol}(E)$ *is dense in $C^k(E)$.*

Proof. Without loss of generality, we can restrict ourselves to the case $E \subset \mathbb{R}$ be open with $K = [0,1] \subset E$ and $k = 1$, where we fix some $f \in C^1(E)$. Moreover, we introduce the difference operator $D_n : C^0([0,1]) \to C^0([0,1])$ defined by

$$(D_n f)(x) = \frac{f\left(\left(1 - \frac{1}{n}\right) x + \frac{1}{n}\right) - f\left(\left(1 - \frac{1}{n}\right) x\right)}{\frac{1}{n}},$$

for $n \in \mathbb{N}_{\geqslant 1}$ and $f \in C^0([0,1])$. Then, it follows for every $x \in [0,1]$ by the mean-value theorem that $D_n f(x) = f'(\xi)$, for some $\xi \in [0,1]$ with $|x - \xi| < 1/n$, such that $D_n f$ converges uniformly to f', as $n \to \infty$.

Now, we conclude that $(B_n f)' = B_{n-1} D_n f$ and $\|B_n\|_{L(C^0([0,1]),C^0([0,1]))} = 1$, and use Weierstrass' result in Theorem 1.33 such that $B_n f$ converges uniformly to f, as $n \to \infty$. In addition, it follows from

$$\begin{aligned} \left\|(B_n f)' - f'\right\|_{C^0([0,1])} &= \left\|B_{n-1}\left(D_n f - f'\right) + B_{n-1} f' - f'\right\|_{C^0([0,1])} \\ &\leqslant \|B_{n-1}\|_{L(C^0([0,1]),C^0([0,1]))} \left\|D_n f - f'\right\|_{C^0([0,1])} \\ &\quad + \left\|B_{n-1} f' - f'\right\|_{C^0([0,1])} \quad \overset{n\to\infty}{\longrightarrow} \quad 0 \end{aligned}$$

that $(B_n f)'$ converges uniformly to f', as $n \to \infty$, which completes the proof. □

Leopoldo Nachbin extended Stone's result in Theorem 1.35 to the function space $C^k(E)$ with an open subset $E \subset \mathbb{R}^d$, or analogously generalized Weierstrass' approximation result with derivatives in Theorem 1.36 by replacing polynomials with more general subalgebras. Hereby, $\nabla f = \left(\frac{\partial f}{\partial x_1}, ..., \frac{\partial f}{\partial x_d}\right)^\top$ denotes the gradient of a function $f \in C^1(E)$.

Theorem 1.37 (Nachbin, [Nac49, p. 1550]). *Let $E \subset \mathbb{R}^d$ be open and assume that $\mathcal{A} \subset C^k(E)$ is a subalgebra. Then, $\mathcal{A}$ is dense in $C^k(E)$ if and only if $\mathcal{A}$ is point separating, vanishes nowhere, and for every $x \in E$ and $v \in \mathbb{R}^d \backslash \{0\}$ there exists some $a \in \mathcal{A}$ such that $v^\top \nabla a(x) \neq 0$.*

Theorem 1.37 can be used to show the approximation of a function $f \in C^k(E)$ by an element of the subalgebra such that for every $\varepsilon > 0$ and compact $K \subset E$ there exists some $a \in \mathcal{A}$ with $\|f - a\|_{C^k(K)} < \varepsilon$, where the function $a \in \mathcal{A}$ approximates not only the function value of $f : E \to \mathbb{R}$, but also its derivatives.

1.2.4 Stochastic Analysis and Pathwise Ito Calculus

In this section, we introduce the foundations of Stochastic Analysis, for which we assume that $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{P})$ denotes a filtered probability space with filtration $\mathbb{F} = (\mathcal{F}_t)_{t \in [0,\infty[}$ satisfying the usual conditions, i.e. $\mathcal{F}_0$ is trivial, $\mathcal{A}$ is $\mathbb{P}$-complete and the filtration is right-continuous, i.e. $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$. For $E \subset \mathbb{R}^d$, an E-valued stochastic process is a measurable map $X : \Omega \times [0,\infty[\to E$, which is called $\mathbb{F}$-*adapted* if the random variable $X_t : \Omega \to E$ is $\mathcal{F}_t$-measurable, for all $t \in [0,\infty[$. We abbreviate the stochastic process by $X = (X(t))_{t \in [0,\infty[}$.

Definition 1.38. *Let X be an $\mathbb{F}$-adapted stochastic process, for which we define the following:*

(i) $X : \Omega \times [0,\infty[\to E$ *is a* (sub-/super-) martingale *if* $\mathbb{E}[|X_i(t)|] < \infty$, *for all* $i = 1, ..., d$ *and* $t \in [0,\infty[$, *and it holds that* $\mathbb{E}[X(t)|\mathcal{F}_s] = X(s)$, *a.s. (resp.* $\geqslant$, *resp.* $\leqslant$*), for all* $i = 1, ..., d$ *and* $0 \leqslant s \leqslant t < \infty$.

(ii) $X : \Omega \times [0,\infty[\to E$ *is a* local (sub-/super-) martingale *if there exists an increasing sequence of stopping times* $(\tau_k)_{k \in \mathbb{N}}$ *with* $\mathbb{P}[\lim_{k\to\infty} \tau_k = \infty] = 1$ *such that the stopped process* X^{τ_k} *defined by* $X^{\tau_k}(t) = X(t \wedge \tau_k)$, *for* $t \in [0,\infty[$, *is a true (sub-/super-) martingale, for all* $k \in \mathbb{N}$.

(iii) $X : \Omega \times [0,\infty[\to E$ *is a* process of finite variation *if for every* $i = 1, ..., d$ *and* $T > 0$, *we have* $\sup_{s,t \in [0,T]} |X_i(t) - X_i(s)| < \infty$, *a.s.*

(iii) $X : \Omega \times [0,\infty[\to E$ *is a* semimartingale *if there exists a local martingale M and a process of finite variation A such that* $X(t) = X(0) + M(t) + A(t)$, *for all* $t \in [0,\infty[$.

An example of a martingale is the one-dimensional Brownian motion $B = (B(t))_{t \in [0,\infty[}$, which is defined by the following properties: B has independent and normally distributed increments such that $B(t) - B(s) \sim \mathcal{N}(0, t - s)$,

B has a.s. continuous sample paths $t \mapsto X(t)$, and it holds that $B(0) = 0$, a.s. Other examples of (local) martingales are the compensated poisson process or other driftless diffusion processes. In addition, we introduce the stochastic integral (Ito integral) of some $\mathbb{F}$-predictable locally integrable process h with respect to the semimartingale X as $L^2(\Omega)$-limit $\int_s^t h(u)dX(u) := \lim_{|\pi_n|\to 0} \sum_{k=1}^n h(t_k)^\top (X(t_{k+1}) - X(t_k))$, for a sequence of partitions $(\pi_n)_{n\in\mathbb{N}}$ of $[s,t]$ with mesh $|\pi_n| := \max_{k=1,...,n} |t_k - t_{k-1}| \to 0$, as $n \to \infty$.

Theorem 1.39 (Ito's formula, [Ito44])**.** *Let X be a semimartingale and assume that $f \in C^2(\mathbb{R}^d)$ is a function. Then, it holds that*

$$f(X(t)) - f(X(0)) = \sum_{i=1}^d \int_0^t \frac{\partial f}{\partial x_i}(X(u))dX_i(u) + \frac{1}{2}\sum_{i,j=1}^d \int_0^t \frac{\partial^2 f}{\partial x_i \partial x_j}(X(u))d\langle X_i, X_j\rangle_u,$$

a.s., for all $t \in [0,\infty[$.

For the applications in this thesis, we introduce the pathwise Ito calculus of Föllmer developed in [Föl81]. We fix a refining sequence of partitions $(\mathbb{T}_n)_{n\in\mathbb{N}}$ of $[0,\infty[$, with $\mathbb{T}_n = \{0 = t_0^n < t_1^n < ...\}$, such that $\mathbb{T}_n \subset \mathbb{T}_{n+1}$ and $t_k^n \to \infty$, as $k \to \infty$. Moreover, we denote the successor of some $t \in \mathbb{T}_n$ by t' defined as $t' = \min\{u \in \mathbb{T}_n;\, u > t\}$ and assume that $|\mathbb{T}_n| := \max_{t\in\mathbb{T}_n} |t' - t| \to 0$, as $n \to \infty$. Hereby, we suppose that the continuous path $X : [0,\infty[\to \mathbb{R}^d$ exhibits pathwise continuous quadratic variation along $(\mathbb{T}_n)_{n\in\mathbb{N}}$ in the sense of [Föl81], where $\mathbb{S}_+^d$ denotes the space of symmetric positive definite matrices in $\mathbb{R}^{d\times d}$

Assumption 1.40. *Suppose that the path $X : [0,\infty[\to \mathbb{R}^d$ admits continuous $\mathbb{S}_+^d$-valued quadratic variation $[X] : [0,\infty[\to \mathbb{S}_+^d$ along $(\mathbb{T}_n)_{n\in\mathbb{N}}$ in the sense of [Föl81], i.e. for any $i,j = 1,...,d$ and $t \in [0,T]$ the sequence*

$$\sum_{s\in\mathbb{T}_n,\, s\leqslant t} \big(X_i(s') - X_i(s)\big)\big(X_j(s') - X_j(s)\big) \tag{1.2.6}$$

converges to a finite limit, as $n \to \infty$, denoted by $[X_i, X_j]_t$, such that $t \mapsto [X_i, X_j]_t$ is continuous.

We note that the limit in (1.2.6) depends on the choice of the partitioning sequence $(\mathbb{T}_n)_{n\in\mathbb{N}}$. However, in the case of sample paths of a semimartingale X, $[X_i, X_j]_t$ coincides with the classical quadratic variation $\langle X_i, X_j\rangle_t$. Under Assumption 1.40, the Ito formula still holds true in this pathwise setting.

Theorem 1.41 (Pathwise Ito formula, [Föl81]). *For an open subset $U \subset \mathbb{R}^d$, let $f \in C^2(U)$ be a function and assume that the path $X : [0, \infty[\to U$ satisfies Assumption 1.40. Then, it follows that*

$$f(X(t)) - f(X(0)) = \int_0^t \nabla f(X(s))^\top dX(s) + \frac{1}{2} \sum_{i,j=1}^{d} \int_0^t D_{ij} f(X(s)) d[X_i, X_j]_s,$$

for all $t \in [0, \infty[$, where the pathwise stochastic integral is defined as

$$\int_0^t \nabla f(X(s))^\top dX(s) = \lim_{n \to \infty} \sum_{s \in \mathbb{T}_n, s \leqslant t} \nabla f(X(s))^\top \left(X(s') - X(s) \right), \qquad (1.2.7)$$

for $t \in [0, \infty[$, and where the series on the right-hand side of (1.2.7) converges absolutely.

Moreover, a function $\xi : [0, \infty[\to U$ is called an *admissible integrand (on U)* if there exists a function $f \in C^2(U)$ with $\xi(t) = \nabla f(X(t))$, for all $t \in [0, \infty[$, where $U \subset \mathbb{R}^d$ is an open subset. In this case, Theorem 1.41 gives an expression for the pathwise stochastic integral, which we call the *Föllmer integral*.

Part I

Universal Approximation

Chapter 2

Universal Approximation in Finite Dimension

Universal approximation on $\mathbb{R}^d$ is a property of a function class, which establishes their density within a given function space of interest. Artificial neural networks are an example of such a class of functions, which were introduced in the seminal work of Warren McCulloch and Walter Pitts in [MP43] and are inspired by the functionality of the human brain. It was only later proven by George Cybenko and Kurt Hornik in [Cyb89] and [Hor91] that this class of functions indeed possesses the universal approximation property.

The proof of [Cyb89] relies on the Stone-Weierstrass theorem and therefore does not provide any approximation rates. The same holds true for the proof of [Hor91], where Fourier methods are additionally applied. Subsequently, the literature on approximation-theoretic results for neural networks with one single hidden layer was rapidly growing so that we only highlight two important results. In [Bar93], the approximation error depends on the number of nodes in the hidden layer, whereas [Can98] showed fundamental lower bounds on approximation rates. The literature on approximation results for neural networks with more than one hidden layer is even richer, and we only mention the work of [BGKP19] on optimal approximation error rates.

Finally, we also introduce recurrent neural networks (RNNs), where we focus on echo-state networks (ESNs) introduced in [Jae10], which is a special class of RNNs, whose universality was shown in [GO18].

2.1 Neural Networks

Inspired by the functionality of a brain, neural networks transmit information through a graph of neurons. In mathematical terms, neural networks are defined as concatenation of affine and non-linear maps.

Definition 2.1. *Let $L, d, N_0, ..., N_L \in \mathbb{N}$ with $L \geqslant 2$, and let $\rho : \mathbb{R} \to \mathbb{R}$ be a function. Let $W_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$ be affine functions given by $W_l(x) = A_l x + b_l$, for some $A_l \in \mathbb{R}^{N_l \times N_{l-1}}$ and $b_l \in \mathbb{R}^{N_l}$. Then, a function $\varphi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ given by*

$$\varphi(x) = (W_L \circ \rho \circ W_{L-1} \circ ... \circ W_2 \circ \rho \circ W_1)(x),$$

for $x \in \mathbb{R}^d$, is called a neural network, *where $(A_l)_{l=1,...,L}$ are the* weight matrices, *$(b_l)_{l=1,...,L}$ the* bias vectors, *and where $\rho : \mathbb{R} \to \mathbb{R}$ denotes the* activation function *acting componentwise. Hereby, L denotes the number of layers (without counting the input layer), $N_1, ..., N_{L-1}$ represent the dimensions of the hidden layers and N_0, N_L of the input and output layer, respectively. The total number of non-zero weights, i.e. the number of non-zero entries in $A_1, ..., A_L$, is denoted by M and called the network's* connectivity.

In this case, we speak of a feed-forward neural network, which can be interpreted as directed graph with nodes arranged in $L + 1$ hierarchical layers and edges only between adjacent layers. The input is fed into the input layer on one side of the graph, whereas the output is returned on the other side, see Figure 2.1.

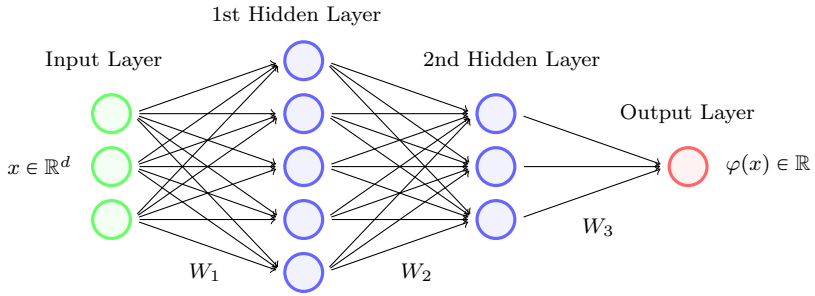


Figure 2.1: An example of a neural network $\varphi : \mathbb{R}^d \to \mathbb{R}$ with two hidden layers, where $d = 3$, $L = 3$ and $(N_1, N_2) = (5, 3)$.

In this chapter, we will always assume that the ouput is one-dimensional, i.e. $N_L = 1$, because the components of the output of some higher dimensional neural networks can be considered separately. In addition, we collect all neural networks $\varphi : \mathbb{R}^d \to \mathbb{R}$ with activation function $\rho : \mathbb{R} \to \mathbb{R}$ and $L - 1$ hidden layers in $\mathcal{NN}^\rho_{L-1}$, where we let $\mathcal{NN}^\rho_{d,M}$ be the set of neural networks with arbitrary hidden layers but connectivity at most M, from which we define the set of arbitrary networks as $\mathcal{NN}^\rho_{d,\infty} := \bigcup_{M \in \mathbb{N}} \mathcal{NN}^\rho_{d,M}$.

Moreover, the activation function $\rho : \mathbb{R} \to \mathbb{R}$ can be chosen differently, where the sigmoid function $\rho(x) = 1/(1 + \exp(-x))$ and the ReLU function $\rho(x) =$

$\max(x, 0)$ appear the most frequently in the literature. Both are continuous, even Lipschitz continuous, whereas only the first one is bounded and smooth. Because of the unboundedness of the ReLU function, some usual approximation arguments cannot be directly applied to neural networks with the ReLU activation function. Besides these two activation functions, there are also other activation functions such as the hyperbolic tangent $\rho(x) = \tanh(x)$, the heaviside function $\rho(x) = \mathbb{1}_{\{x \geqslant 0\}}$, and the softplus function $\rho(x) = \log(1 + \exp(x))$.

2.2 Universal Approximation

Neural networks have been successfully applied for various problems in different areas because of their approximation capabilities. In mathematical terms, this property of neural networks is usually shown in so-called universal approximation theorems, which establishes density of neural networks within a function space of interest. Following the ideas of [Cyb89], we show universal approximation of neural networks $\varphi \in \mathcal{NN}_1^\rho$ with one hidden layer, which can be easily generalized to deep neural networks with finitely many hidden layers. For this purpose, we introduce the following two properties of an activation function.

Definition 2.2. *Let $K \subset \mathbb{R}^d$ be compact. The function $\rho : \mathbb{R} \to \mathbb{R}$ is called* discriminatory (on K) *if for every finite signed Radon measure $\mu : \mathcal{F} \to \mathbb{R}$ satisfying for every $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that*

$$\int_K \rho\left(a^\top x + b\right) d\mu(x) = 0, \tag{2.2.1}$$

it follows that $\mu : \mathcal{F} \to \mathbb{R}$ vanishes everywhere. Hereby, $\mathcal{F}$ denotes the Borel σ-algebra of K.

Definition 2.3. *The function $\rho : \mathbb{R} \to \mathbb{R}$ is called* sigmoidal *if $\lim_{x \to -\infty} \rho(x) \neq \lim_{x \to \infty} \rho(x)$, where both limits are assumed to exist, and for every $\varepsilon > 0$ there exists some $x_0 \in \mathbb{R}$ such that $|\rho(x_0)| < \varepsilon$.*

The condition for an activation function $\rho : \mathbb{R} \to \mathbb{R}$ to be discriminatory is difficult to verify, whereas the sigmoidal property is characterised by the easier condition that the limits in opposite directions must differ. Following [Cyb89, Lemma 1], we show that the sigmoidal condition implies the discriminatory property.

Lemma 2.4. *Let $\rho : \mathbb{R} \to \mathbb{R}$ be bounded, Borel-measurable and sigmoidal. Then, ρ is discriminatory.*

Proof. Let $\rho : \mathbb{R} \to \mathbb{R}$ be a sigmoidal function with $\rho_- := \lim_{x\to-\infty} \rho(x) \neq \lim_{x\to\infty} \rho(x) =: \rho_+$ and suppose that $\mu : \mathcal{F} \to \mathbb{R}$ is a finite Radon measure satsfying (2.2.1). Fix $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$, for which we observe that

$$\rho\left(\lambda\left(a^\top x + b\right) + \theta\right) \quad \overset{\lambda\to\infty}{\longrightarrow} \quad \begin{cases} \rho_+ & a^\top x + b > 0 \\ \rho(\theta) & a^\top x + b = 0 \\ \rho_- & a^\top x + b < 0 \end{cases},$$

for all $\theta \in \mathbb{R}$. Because $\rho : \mathbb{R} \to \mathbb{R}$ is bounded, we can apply dominated convergence to conclude that

$$\begin{aligned} 0 = \int_K \rho\left(\lambda\left(a^\top x + b\right) + \theta\right) d\mu(x) \quad &\overset{\lambda\to\infty}{\longrightarrow} \quad \rho_- \int_K \mathbb{1}_{\{a^\top x + b < 0\}} d\mu(x) \\ &+ \rho_+ \int_K \mathbb{1}_{\{a^\top x + b > 0\}} d\mu(x), \end{aligned}$$

where we have used that $\theta \in \mathbb{R}$ can be chosen arbitrarily, which implies that $\rho(\theta)\mu\left(\left\{a^\top x + b = 0\right\}\right) = 0$. Because every continuous function $h \in C^0(\mathbb{R})$ can be approximated pointwise by step functions, i.e. by linear combinations of $\mathbb{1}_{\{z+b<0\}}$ and $\mathbb{1}_{\{z+b>0\}}$ using $\rho_- \neq \rho_+$, we conclude by dominated convergence that

$$\int_K g\left(a^\top x\right) d\mu(x) = 0,$$

for all $a \in \mathbb{R}^d$ and continuous functions $g \in C^0(\mathbb{R})$. Choosing $g(x) = \exp(x)$, we observe for every $n \in \mathbb{N}_{\geqslant 1}$ and $a_1, ..., a_n \in \mathbb{R}^d$ that

$$\int_K e^{a_1^\top x} \cdots e^{a_n^\top x} d\mu(x) = \int_K e^{(a_1 + ... + a_n)^\top x} d\mu(x) = 0. \tag{2.2.2}$$

Let $\mathcal{A} = \operatorname{span}\left\{x \mapsto e^{a_1^\top x} \cdots e^{a_n^\top x};\ n \in \mathbb{N}_{\geqslant 1},\ a_k \in \mathbb{R}^d\right\}$ such that by taking finite linear combinations of (2.2.2), we have

$$\int_K a(x) d\mu(x) = 0, \tag{2.2.3}$$

for all $a \in \mathcal{A}$. Since $\mathcal{A}$ is a point separating subalgebra, which vanishes nowhere, it follows by the Stone-Weierstrass in Theorem 1.35 that $\mathcal{A}$ is dense in $C^0(K)$, which shows that (2.2.3) holds true, for all $f \in C^0(K)$. This however implies that $\mu : \mathcal{F} \to \mathbb{R}$ vanishes everywhere, whence $\rho : \mathbb{R} \to \mathbb{R}$ is discriminatory. □

Theorem 2.5 (Universal Approximation, [Cyb89, Theorem 1]). *Let $K \subset \mathbb{R}^d$ be compact and assume that $\rho : \mathbb{R} \to \mathbb{R}$ is continuous and discriminatory. Then, $\mathcal{NN}_1^\rho$ is dense in $C^0(K)$, i.e. for every $f \in C^0(K)$ and $\varepsilon > 0$ there exists some $\varphi \in \mathcal{NN}_1^\rho$ such that*

$$\|f - \varphi\|_{C^0(K)} = \sup_{x \in K} |f(x) - \varphi(x)| < \varepsilon.$$

Proof. Assume by contradiction that $\mathcal{NN}_1^\rho$ is not dense in $C^0(K)$. Then, there exists by Hahn-Banach in Corollary 1.14 a non-zero, continuous linear functional $l : C^0(K) \to \mathbb{R}$ vanishing on $\mathcal{NN}_1^\rho$, which can be represented by Riesz representation with some finite signed Radon measure $\mu : \mathcal{F} \to \mathbb{R}$ such that $l(f) = \int_K f(x) d\mu(x)$, for all $f \in C^0(K)$. However, because $\rho : \mathbb{R} \to \mathbb{R}$ is discriminatory, it follows that $\mu : \mathcal{F} \to \mathbb{R}$ vanishes everywhere, which contradicts the assumption that $l : C^0(K) \to \mathbb{R}$ is non-zero. □

Discriminatory functions are not the only functions, for which $\mathcal{NN}_1^\rho$ is dense in $C^0(K)$. For example, if $\rho \in L^1(\mathbb{R})$ is a function with $\int_{\mathbb{R}} \rho(x) dx \neq 0$, it follows by Wiener's Tauberian theorem in [Wie32] that $\mathcal{NN}_1^\rho$ is dense in $L^1(\mathbb{R})$. On the other hand, [Hor91, Theorem 2] and [Pin99, Theorem 3.1] show that $\mathcal{NN}_1^\rho$ is also dense in $C^0(K)$, if $\rho : \mathbb{R} \to \mathbb{R}$ is bounded, continuous and non-constant or if $\rho : \mathbb{R} \to \mathbb{R}$ is continuous and non-polynomial, which includes in particular the ReLU activation function $\rho(x) = \max(x, 0)$.

Universal approximation for deep neural networks with more than one hidden layer follows by iteration. Since $\mathcal{NN}_1^\rho$ is an additively closed and point separating family, the arguments in Lemma 2.4 can be adapted to show universal approximation of neural networks with two hidden layers. By iterating the argument, we obtain universal approximation for neural networks with finitely many hidden layers.

In addition, Theorem 2.5 can be generalized to L^p-spaces on measure spaces (Ω, Σ, ν), to Besov spaces $B^s_{p,q}(\mathbb{R}^d)$ and differentiable function spaces $C^k(E)$, where the latter is considered in the next section. The essential ingredients are an additively closed and point separating set of functions, which is here given by the set of linear maps $\{x \mapsto a^\top x;\, a \in \mathbb{R}^d\}$, a weak topology, which allows to apply the dominated convergence theorem, and a discriminatory activation function for the corresponding linear functionals.

2.3 Universal Approximation with Derivatives

In this section, we consider the approximation of a differentiable function by neural networks, where not only the function itself but also its derivatives are

simultaneously approximated. This could be of particular interest for applications in Physics, where the velocity of a path needs to be approximated in addition. More precisely, we try to find a neural network with smooth activation function such that $\|f-\varphi\|_{C^k(K)}<\varepsilon$, for some given $\varepsilon>0$ and compact, non-empty subset $K\subset\mathbb{R}^d$. A first approach was provided in [HSW90] using Fourier arguments, whereas the convolution idea of [Hor91] is presented in the following.

Theorem 2.6 (Universal Approximation on $C^k(E)$, [Hor91, Theorem 3])**.** *Let $E\subset\mathbb{R}^d$ be open and assume that $\rho:\mathbb{R}\to\mathbb{R}$ is k-times differentiable and discriminatory. Then, $\mathcal{NN}_1^\rho$ is dense in $C^k(E)$, i.e. for every $f\in C^k(E)$, $\varepsilon>0$ and compact subset $K\subset E$ there exists some $\varphi\in\mathcal{NN}_1^\rho$ such that*

$$\|f-\varphi\|_{C^k(K)}=\sum_{\alpha\in\mathbb{N}_k^d}\sup_{x\in K}|\partial^\alpha f(x)-\partial^\alpha\varphi(x)|<\varepsilon.$$

Proof. Assume by contradiction that $\mathcal{NN}_1^\rho$ is not dense in $C^k(E)$. Then, there exists by Hahn-Banach in Corollary 1.14 a non-zero, continuous linear functional $l:C^k(E)\to\mathbb{R}$ vanishing on $\mathcal{NN}_1^\rho$, which can be represented by Riesz representation with finite signed Radon measures $(\mu_\alpha)_{\alpha\in\mathbb{N}_k^d}:\mathcal{F}\to\mathbb{R}$ such that $l(f)=\sum_{\alpha\in\mathbb{N}_k^d}\int_{\mathbb{R}^d}\partial^\alpha f(x)d\mu_\alpha(x)$, for all $f\in C^k(\mathbb{R}^d)$. Let $\eta:\mathbb{R}^d\to\mathbb{R}$ be the bump function defined by

$$\eta(x)=\begin{cases}Ce^{-\frac{1}{1-\|x\|^2}} & \|x\|<1\\ 0 & \text{otherwise}\end{cases},$$

for $x\in\mathbb{R}^d$, where $C>0$ is chosen such that $\int_{\mathbb{R}^d}\eta(x)dx=1$. For $\varepsilon>0$, we define $\eta_\varepsilon:\mathbb{R}^d\to\mathbb{R}$ by $\eta_\varepsilon(x)=\frac{1}{\varepsilon^d}\eta\left(\frac{x}{\varepsilon}\right)$, for $x\in\mathbb{R}^d$. If $f:\mathbb{R}^d\to\mathbb{R}$ is locally integrable, we define the convolution $\eta_\varepsilon*f\in C^\infty(\mathbb{R}^d)$ by

$$(\eta_\varepsilon*f)(x)=\int_{\mathbb{R}^d}\eta_\varepsilon(x-y)f(y)dy,$$

for $x\in\mathbb{R}^d$, such that $\partial^\alpha(\eta_\varepsilon*f)(x)=\left(\partial^\alpha\eta_\varepsilon*f\right)(x)$, for all $\alpha\in\mathbb{N}_k^d$ and $x\in\mathbb{R}^d$. In addition, if $f:\mathbb{R}^d\to\mathbb{R}$ is bounded, it holds that $\|\eta_\varepsilon*f\|_{C^0(\mathbb{R}^d)}\leqslant\|f\|_{C^0(\mathbb{R}^d)}$, and if $f:\mathbb{R}^d\to\mathbb{R}$ is continuous, $\eta_\varepsilon*f$ converges uniformly on compacta to f, as $\varepsilon\to0$. Similarly for $(\mu_\alpha)_{\alpha\in\mathbb{N}_k^d}:\mathcal{F}\to\mathbb{R}$, we define the convolution $\eta_\varepsilon*\mu\in C^\infty(\mathbb{R}^d)$ by

$$(\eta_\varepsilon*\mu_\alpha)(x)=\int_{\mathbb{R}^d}\eta_\varepsilon(x-y)d\mu_\alpha(y),$$

for $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{N}_k^d$. Now, we observe that for every continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and $\alpha \in \mathbb{N}_k^d$, it follows by Fubini's theorem that

$$\begin{aligned}\int_{\mathbb{R}^d} f(x)(\eta_\varepsilon * \mu_\alpha)(x)dx &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} f(x+y)d\mu_\alpha(x) \right) \eta_\varepsilon(y)dy \\ &= \int_{\mathbb{R}^d} (\eta_\varepsilon * f)(x)d\mu_\alpha(x),\end{aligned} \tag{2.3.1}$$

such that $\lim_{\varepsilon\to 0} \int_{\mathbb{R}^d} f(x)(\eta_\varepsilon * \mu_\alpha)(x)dx = \int_{\mathbb{R}^d} f(x)d\mu_\alpha(x)$. Hence, we define the continuous linear functional $l_\varepsilon : C^k(\mathbb{R}^d) \to \mathbb{R}$ by

$$l_\varepsilon(f) = \sum_{\alpha\in\mathbb{N}_k^d} \int_{\mathbb{R}^d} \partial^\alpha f(x)(\eta_\varepsilon * \mu_\alpha)(x)dx,$$

for $f \in C^k(\mathbb{R}^d)$, which is the mollification of $l : C^k(E) \to \mathbb{R}$. Using (2.3.1) and integration by parts, we have

$$\begin{aligned}l_\varepsilon(f) &= \int_{\mathbb{R}^d} \left(\sum_{\alpha\in\mathbb{N}_k^d} \int_{\mathbb{R}^d} \partial^\alpha f(x+y)d\mu_\alpha(x) \right) \eta_\varepsilon(y)dy = \int_{\mathbb{R}^d} l(f(\cdot+y))\eta_\varepsilon(y)dy \\ &= \int_{\mathbb{R}^d} \underbrace{\left(\sum_{\alpha\in\mathbb{N}_k^d} (-1)^{|\alpha|} \partial^\alpha (\eta_\varepsilon * \mu_\alpha)(x) \right)}_{=:h_\varepsilon(x)} f(x)dx,\end{aligned}$$

for all $f \in C^k(\mathbb{R}^d)$, where we define $h_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ as above. Since $l : C^k(E) \to \mathbb{R}^d$ vanishes on $\mathcal{NN}_1^\rho$, it follows that $l(\varphi(\cdot + y)) = 0$, for all $\varphi \in \mathcal{NN}_1^\rho$ and $y \in \mathbb{R}^d$. From this, we conclude that $l_\varepsilon(\varphi) = 0$, for all $\varphi \in \mathcal{NN}_1^\rho$ and by using that $\rho : \mathbb{R} \to \mathbb{R}$ is discriminatory for the measure $h_\varepsilon(x)dx$, it follows that $h_\varepsilon(x) = 0$, for all $x \in \mathbb{R}^d$. Therefore, $l_\varepsilon : C^k(\mathbb{R}^d) \to \mathbb{R}$ vanishes everywhere and using that $l(f) = \lim_{\varepsilon\to 0} l_\varepsilon(f)$, for all $f \in C^k(E)$, it follows that $l : C^k(E) \to \mathbb{R}$ vanishes also everywhere, which, however, contradicts the assumption that $l : C^k(E) \to \mathbb{R}$ is non-zero. Hence, $\mathcal{NN}_1^\rho$ is dense in $C^k(E)$. □

Universal approximation with derivatives of deep neural networks can be obtained by iterating the argument, which shows that neural networks with finitely many hidden layers and smooth activation function can be used to approximate differentiable functions including their derivatives arbitrarily well.

2.4 Recurrent Neural Networks

Recurrent neural networks (RNNs) can be considered as neural networks with connections along a temporal sequence, such that RNNs are able to exhibit temporally dynamic behaviour. RNNs are used in handwriting and speech recognition, for which there exist various architectures, ranging from the first RNN, called the Hopfield RNN in [Hop82], to long short-term memory networks (LSTMs) in [HS97], and echo-state networks (ESNs) in [Jae10] and [GO18]. In the following, we consider the latter class of RNNs.

Echo-State networks (ESNs) are a special class of recurrent neural networks, where the neurons are initialized randomly and only the linear readout is trained. With this dynamical reservoir, echo-state networks are rich enough to approximate dynamical systems arbitrarily well, where a dynamical system $F = (g, h)$ is defined as

$$\begin{cases} u_t &= g(u_{t-1}, x_t) \\ y_t &= h(u_t) \end{cases},$$

for $t \in \mathbb{Z}$, where $(x_t)_{t\in\mathbb{Z}}$ is the input sequence in $\mathbb{R}^d$, $(u_t)_{t\in\mathbb{Z}}$ is the hidden-state sequence in $\mathbb{R}^n$, and $(y_t)_{t\in\mathbb{Z}}$ is the output sequence in $\mathbb{R}^m$. Hereby, $g : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are functions and we say that $F = (g, h)$ has the *echo-state property (ESP)* if the system $F = (g, h)$ has the existence of solutions property and the solutions are unique. Then, an echo-state network is a dynamical system of a special form.

Definition 2.7. *Let $\rho : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function. Then, for an input sequence $(x_t)_{t\in\mathbb{Z}}$, the system $\varphi = (A, C, W, b, \rho)$ defined by*

$$\begin{cases} u_t &= \rho(Au_{t-1} + Cx_t + b) \\ y_t &= Wu_t \end{cases},$$

for $t \in \mathbb{Z}$, is called an echo-state network (ESN)*, where $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{n\times d}$, $W \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$. Hereby, $(u_t)_{t\in\mathbb{Z}}$ is called the* hidden-state sequence *and the* output sequence *is denoted by $(y_t)_{t\in\mathbb{Z}}$.*

Let $D \subset \mathbb{R}^d$ be the domain of the input sequences, and define $D^{\mathbb{Z}} = \{x = (..., x_{-1}, x_0, x_1, ...);\ x_t \in D\}$, $D^{\mathbb{Z}_-} = \{x = (..., x_{-1}, x_0);\ x_t \in D\}$ as well as $D^{\mathbb{Z}_+} = \{x = (x_0, x_1, ...);\ x_t \in D\}$ as the (semi-)infinite sequences with values in D, which are equipped with the supremum norm $\|x\|_\infty = \sup_{t\in\mathbb{Z}} \|x_t\|$, for $x \in D^{\mathbb{Z}}$, where the supremum is taken over $t \in \mathbb{Z}_-$ (resp. $t \in \mathbb{Z}_+$) for $D^{\mathbb{Z}_-}$ (resp. $D^{\mathbb{Z}_+}$). Let $E \subset \mathbb{R}^N$ be the domain of the hidden-state sequence, for which we introduce $E^{\mathbb{Z}}$, $E^{\mathbb{Z}_-}$ and $E^{\mathbb{Z}_+}$ analogously.

We call the map $U : D^{\mathbb{Z}_-} \to E^{\mathbb{Z}_-}$ a *filter*, whereas $H : D^{\mathbb{Z}_-} \to E$ is a *functional*. A filter $U : D^{\mathbb{Z}_-} \to E^{\mathbb{Z}_-}$ is called *causal* if for every $x, y \in D^{\mathbb{Z}_-}$ satisfying $x_s = y_s$, for all $s \leqslant t$, for some $t \in \mathbb{Z}_-$, it follows that $U(x)_t = U(y)_t$, whereas the filter $U : D^{\mathbb{Z}_-} \to E^{\mathbb{Z}_-}$ is called *time-invariant* if $U(x)_{t-s} = U(x_{\_-s})_t$, for all $s, t \in \mathbb{Z}$ with $t \leqslant s$. Moreover, a causal and time-invariant filter $U : D^{\mathbb{Z}_-} \to E^{\mathbb{Z}_-}$ satisfies the *fading memory property (FMP)* if it is continuous for the weighted sequence topologies, see [GO18, Definition 2.5]. In addition, we can associate to every filter $U : D^{\mathbb{Z}_-} \to E^{\mathbb{Z}_-}$ a functional $H_U : D^{\mathbb{Z}_-} \to \mathbb{R}^n$ by defining $H_U(x) = U(x)_0$, for $x \in D^{\mathbb{Z}_-}$.

An echo-state network $\varphi = (A, C, W, b, \rho)$ can be seen as a filter sending the input $(x_t)_{t\in\mathbb{Z}_-}$ to the output $(y_t)_{t\in\mathbb{Z}_-}$, which we denote by a slight abuse of notation with $y = \varphi(x)$. In this case, $\varphi : D^{\mathbb{Z}_-} \to (\mathbb{R}^d)^{\mathbb{Z}_-}$ has by [GO18, Corollary 3.2] the *echo-state property* and the *fading memory property* if $\rho : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L_ρ such that $\|A\|_2 L_\rho = \sigma_{\max}(A) L_\rho < 1$, where $\sigma_{\max}(A)$ denotes the largest singular value of the matrix $A \in \mathbb{R}^{n\times n}$. In addition, let $B_1^d(0)$ be the open ball of radius one around the origin in $\mathbb{R}^d$ and call $\rho : \mathbb{R} \to \mathbb{R}$ a *squashing function* if $\lim_{x\to-\infty} \rho(x) = -1$ and $\lim_{x\to\infty} \rho(x) = 1$, e.g. the shifted sigmoid function $\rho(x) = 2/(1 + \exp(-x)) - 1$ or the tangens hyperbolicus $\rho(x) = \tanh(x)$.

Theorem 2.8 (Universal Approximation of ESN, [GO18, Theorem 4.1])**.** *Let $U : B_1^d(0)^{\mathbb{Z}_-} \to (\mathbb{R}^d)^{\mathbb{Z}_-}$ be a causal and time-invariant filter, which has the fading memory property, and let $\rho : \mathbb{R} \to \mathbb{R}$ be a continuous, non-decreasing squashing function. Then, for every $\varepsilon > 0$ and weighting sequence $w \in \mathbb{Z}_-$ there exists an echo-state network $\varphi = (A, C, W, b, \rho)$ such that*

$$\|U(x) - \varphi(x)\|_\infty = \sup_{t\in\mathbb{Z}_-} \|U(x)_t - \varphi(x)_t\| < \varepsilon, \tag{2.4.1}$$

for all input sequences $(x_t)_{t\in\mathbb{Z}_-}$ with $x_t \in B_1^d(0)$. If the echo-state network $\varphi = (A, C, W, b, \rho)$ satisfies additionally the echo-state property, then the filter $\varphi : B_1^d(0)^{\mathbb{Z}_-} \to (\mathbb{R}^d)^{\mathbb{Z}_-}$ is unique and time-invariant. In addition, the corresponding reservoir functional $H_\varphi : B_1^d(0)^{\mathbb{Z}_-} \to \mathbb{R}^d$ satisfies

$$\|H_U(x) - H_\varphi(x)\| < \varepsilon,$$

for all input sequences $(x_t)_{t\in\mathbb{Z}_-}$ with $x_t \in B_1^d(0)$.

In applications, the weight matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times d}$ and the bias $b \in \mathbb{R}^n$ of an echo-state network are randomly drawn from a given distribution, and only the linear readout $W \in \mathbb{R}^{m\times n}$ is trained. Choosing the mean squared error as loss functions, the training effort reduces to a simple linear regression, which can be computed efficiently also for high dimensional problems.

Chapter 3

Universal Approximation on Weighted Space

The universal approximation property of neural networks was proven for different families of functions over finite dimensional spaces, as described in the previous chapter. However, in many cases these classical results cannot be directly extended to infinite dimensional spaces, which are needed, e.g. for applications in functional data analysis, dynamical systems and partial differential equations (PDEs). For this purpose, we introduce so-called *abstract neural networks* (ANNs) between infinite dimensional spaces, which are defined as the infinite dimensional analogue of neural networks between Euclidean spaces, where we use an additive family as affine functions and apply a non-linear transformation on each hidden layer. Similar generalizations of neural networks to infinite dimensional spaces were recently obtained in [KB20] by inserting finite dimensional spaces in between and in [BDG21] for functionals defined on a Fréchet space.

In order to show the universal approximation property of abstract neural networks, we rely on classical approximation theory including the Stone-Weierstrass theorem and results of Nachbin type. Inspired by Leopoldo Nachbin's weighted approximation problem in [Nac65], we use the structure of a weighted space to consider non-compact input spaces, where a weight function controls the approximation outside of large compact sets and the growth of the functions to be approximated is controlled by the growth of the weight function. For this purpose, we extend the vector-valued Stone-Weierstrass of R. Creighton Buck in [Buc58] to this weighted setting, where the measure theoretic extension of the Stone-Weierstrass theorem of Erret Bishop in [Bis61] should also be mentioned as well as the quantitative version in [Mac77], known as the Lemma of Machado, reducing the approximation problem to antisymmetric subsets.

Subsequently, we show that abstract neural networks with smooth activation functions are not only able to approximate the values of a particular function of interest, but also its derivatives. For this purpose, we generalize Nachbin's result for differentiable functions in [Nac49] to this weighted setting, where we use convenient analysis introduced by Andreas Kriegl and Peter W. Michor in [KM97] to define differentiable calculus on infinite dimensional manifolds. Hereby, we apply several ideas of the classical approximation theory for differentiable functions on infinite dimensional spaces, for which [Lla86] gives a useful overview, where the Nachbin type results of João B. Prolla in [PG76] and [AP80] should also be mentioned.

Finally, we provide two particular applications of the universal approximation results for (smooth) abstract neural networks, which can be unified in this vector-valued framework of a weighted spaces. First, we show that the inner weights of an abstract neural network can be initialized randomly and only the linear readout needs to be trained in order to get a sufficient approximation of a given function. Secondly, we present an application related to path space functionals, which can be learned using the signature of a path, relying on the theory of rough paths introduced by Terry Lyons in [LCL07].

3.1 Motivational Examples between Infinite Dimensional Spaces

In order to obtain an overview of the applicability of the results, we first consider three particular examples. The first example is within functional data analysis, which is a branch of statistics trying to analyse data, where every sample is considered as a continuous function. Hence, one would like to learn a functional $f : C^0(K) \to \mathbb{R}$, where $K \subset \mathbb{R}^n$ is a compact subset. Then, the results of this chapter show that the functional $f : C^0(K) \to \mathbb{R}$ can be approximated by learning an abstract neural network $\varphi : C^0(K) \to \mathbb{R}$ of the form

$$\varphi(x) = \sum_{k=1}^{N} \beta_k \rho \left(\int_K x(u) d\nu_k(u) + b_k \right),$$

for $x \in C^0(K)$, where $N \in \mathbb{N}_{\geqslant 1}$, $\nu_1, ..., \nu_N$ are finite signed Radon measures on K, $\beta_k, b_k \in \mathbb{R}$, and $\rho : \mathbb{R} \to \mathbb{R}$ denotes an activation function. The neural network is still "abstract" because it consists of mathematically abstract objects such as Radon measures. By replacing them for example by conventional neural networks playing the role of a density function, the abstract neural network can be implemented on a computer.

In the second example, we would like to learn a non-anticipative relation between two paths. More precisely, let $x \in C^0([0,T];\mathbb{R}^n)$ be an input path and

let $y \in C^0([0,T];\mathbb{R})$ be the output path given by $y(t) = f(t,x)$, for $t \in [0,T]$, such that y only depends on x up to time t. For the stopped path x^t defined by $x^t(s) = x(s \wedge t)$, for $s \in [0,T]$, we introduce the *space of stopped paths* $\Lambda_T = \left\{(t,x^t);\ (t,x) \in [0,T] \times C^0([0,T],\mathbb{R}^n)\right\}$ as in [Dup09] and [CF10], which defines together with $d_\infty((t,x),(s,y)) = |t-s| + \sup_{u\in[0,T]} \|x^t(u) - y^s(u)\|$ a complete metric space. Hence, a measurable map $f : \Lambda_T \to \mathbb{R}$ satisfies $f(t,x) = f(t,x^t)$, for all $(t,x) \in \Lambda_T$, which shows the non-anticipative behaviour of $f : \Lambda_T \to \mathbb{R}$. In this setting, we approximate a non-anticipative functional $f : \Lambda_T \to \mathbb{R}$ by an abstract neural network $\varphi : \Lambda_T \to \mathbb{R}$ of the form

$$\varphi(t,x) = \sum_{k=1}^{N} \beta_k \rho \left(a_k t + \sum_{i=1}^{n} \int_0^t x_i(u) d\nu_{ik}(u) + b_k \right),$$

for $(t,x) \in \Lambda_T$, where ν_{ik} are Radon measures, $\beta_k, a_k, b_k \in \mathbb{R}$, and $\rho : \mathbb{R} \to \mathbb{R}$ is an activation function. In this case, the Radon measures can also be replaced by conventional neural networks for the implementation.

Finally, we try to learn the solution operator of a partial differential equation $\partial_t x(t,u) + A(x(t,\cdot))(u) = 0$, where $A : W^{k,p}(\mathbb{R}^n) \to W^{k,p}(\mathbb{R}^n)$ is a non-linear operator on the Sobolev space $W^{k,p}(\mathbb{R}^n)$, see [Rau91], [AF08], [Bré11] and [Eva15]. We approximate the solution operator $W^{k,p}(\mathbb{R}^n) \to W^{k,p}(\mathbb{R}^n)$ propagating the initial solution $x_0 = x(0,\cdot) \in W^{k,p}(\mathbb{R}^n)$ to the solution $x(t,\cdot) \in W^{k,p}(\mathbb{R}^n)$ at some fixed time $t > 0$, by learning an abstract neural network $\varphi : W^{k,p}(\mathbb{R}^n) \to W^{k,p}(\mathbb{R}^n)$ of the form

$$\varphi(x_0) = \sum_{k=1}^{N} \beta_k \rho\left(a_k * x_0 + b_k\right), \quad \text{for } x_0 \in W^{k,p}(\mathbb{R}^n),$$

with $(a_k * x_0)(v) := \int_{\mathbb{R}^n} a_k(u) x_0(v-u) du$, where $\beta_k \in \mathbb{R}$, $a_k \in C_c^\infty(\mathbb{R}^n)$ and $b_k \in W^{k,p}(\mathbb{R}^n)$. Hereby, the convolution $a_k * x_0 \in W^{k,p}(\mathbb{R}^n)$ is well-defined by similar arguments as in [Bré11, Lemma 9.1] for the case $k = 1$, and $\rho : W^{k,p}(\mathbb{R}^n) \to W^{k,p}(\mathbb{R}^n)$ is a non-linear operator. For the implementation, the functions $a_k \in C_c^\infty(\mathbb{R}^n)$ and $b_k \in W^{k,p}(\mathbb{R}^n)$ might be replaced by conventional neural networks in $\mathcal{NN}_{n,\infty}^{\rho_0}$ with $\rho_0 \in C_c^\infty(\mathbb{R})$ and $\rho_0 \in W^{k,p}(\mathbb{R})$, respectively.

These three examples provide a first overview over the broad range of applications of this approach. From a mathematical perspective, the input space is allowed to violate the vector space structure (second example), whereas the output space can also be infinite dimensional (third example).

For more background in Functional Analysis, we refer the reader to the textbooks [KM97] and [SW99], whereas the textbooks [Pro77] and [Lla86] give useful overviews on the vector-valued approximation using polynomial algebras and the approximation of differentiable functions including the derivatives.

3.2 Weighted Space and Stone-Weierstrass

For the universal approximation result between infinite dimensional spaces, we use the structure of a weighted space for the input and hidden layer space. Weighted spaces were first introduced in [RS06] for Kolmogorov equations in infinite dimensions, and also used in [DT10] and [CT19] as well as [CT20] in the context of splitting schemes for (stochastic) partial differential equations ((S)PDEs) and generalized Feller processes.

3.2.1 Functions on Weighted Space

In the following, the input space (X, τ_X) is assumed to be a completely regular Hausdorff topological space, whereas the output space (Y, τ_Y) is supposed to be a complete locally convex topological vector space whose topology τ_Y is generated by the family of continuous seminorms $\mathcal{P} = \{p : Y \to \mathbb{R} \text{ seminorm}\}$.

Definition 3.1. *We call $\psi : X \to]0, \infty[$ an* admissible weight function (on (X, τ_X)) *if every preimage $K_R := \psi^{-1}(]0, R]) = \{x \in X;\, \psi(x) \leqslant R\}$ is compact with respect to τ_X, for all $R > 0$.*

An admissible weight function $\psi : X \to]0, \infty[$ is necessarily lower semicontinuous and bounded from below by a strictly positive constant. In this case, the pair (X, ψ) is called a *weighted space*, which is by definition σ-compact with respect to τ_X because we have $X = \bigcup_{R \in \mathbb{N}} K_R$. If (X, τ_X) is furthermore a separable locally convex topological vector space and $\psi : X \to]0, \infty[$ is convex, then $\psi : X \to]0, \infty[$ is continuous on $E \subset X$ if and only if E is locally compact with respect to τ_X, see [CT20, Remark 2.2].

In the following, we present some particular examples of weighted spaces (X, ψ), where the compactness of the preimages $K_R = \psi^{-1}(]0, R])$ needs to be verified with a suitable criterion (e.g. Arzelà-Ascoli type theorems as in Theorem 1.16). Usually, one does not recover the full space of interest in this σ-compact setting, but a large subspace of the input space. In particular, we give some examples related to the input spaces of the abstract neural networks used in the introductory examples in Section 3.1.

Example 3.2. *In the following, we give particular examples of admissible weight functions $\psi : X \to]0, \infty[$.*

(i) For the Euclidean space $X = \mathbb{R}^n$ with the topology induced by the norm $\|x\| = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$, we choose the weight function $\psi(x) = \exp(\|x\|)$ such that $K_R = \psi^{-1}(]0, R])$ is bounded and closed, and thus compact in $\mathbb{R}^n$, for all $R > 0$. Hence, $(\mathbb{R}^n, \psi)$ is a weighted space.

(ii) If X is defined as the dual of a Banach space $(S, \|\cdot\|_S)$ equipped with the weak-$*$-topology, the weight function can be defined as $\psi(x) = \exp(\|x\|_{S*})$ such that $K_R = \psi^{-1}(]0, R])$ is by Banach-Alaoglu in Theorem 1.15 compact in the weak-$*$-topology, for all $R > 0$. Hence, (S^*, ψ) is a weighted space.

(iii) For some $\alpha > 0$ and compact subset $\Omega \subset \mathbb{R}^n$, let $X = C^\alpha(\Omega)$ be the space of α-Hölder continuous functions $x : \Omega \to \mathbb{R}$ such that the Hölder norm $\|x\|_{C^\alpha(\Omega)} := \|x\|_{C^0(\Omega)} + \|x\|_\alpha$ is finite, where

$$\|x\|_\alpha = \sup_{\substack{u,v \in \Omega \\ u \neq v}} \frac{|x(u) - x(v)|}{\|u - v\|^\alpha},$$

for $x \in C^\alpha(\Omega)$. Then, $C^\alpha(\Omega)$ is the dual of a Banach space $(S, \|\cdot\|_S)$ (see Section 3.6) such that $C^\alpha(\Omega)$ can be equipped with a weak-$*$-topology. In this case, the weight function $\psi(x) = \exp\left(\|x\|_{C^\alpha(\Omega)}\right)$ is by Banach-Alaoglu in Theorem 1.15 admissible and $(C^\alpha(\Omega), \psi)$ is a weighted space.

(iv) Instead of the weak-$*$-topology in (iii), we could consider $X = C^\alpha(\Omega)$ with the norm topology generated by $\|\cdot\|_{C^0(\Omega)}$ such that $(X, \|\cdot\|_{C^0(\Omega)})$ is an incomplete normed vector space. In this case, we define the weight function $\psi(x) = \exp\left(\|x\|_{C^\alpha(\Omega)}\right)$ such that $K_R = \psi^{-1}(]0, R])$ is by Arzelà-Ascoli in Corollary 1.17 compact with respect to $\|\cdot\|_{C^0(\Omega)}$, for all $R > 0$. Hence, $(C^\alpha(\Omega), \psi)$ is a weighted space.

(v) For some $\alpha > 0$ and $T > 0$, let $X = \Lambda_T^\alpha$ be the space of stopped α-Hölder continuous paths with metric $d_\infty : \Lambda_T^\alpha \times \Lambda_T^\alpha \to [0, \infty[$ introduced in Section 3.1, which can also be expressed as quotient space $\Lambda_T^\alpha = [0, T] \times C^\alpha([0, T]; \mathbb{R}^n)/\sim$ by identifying $(t, x) \sim (s, y)$ if and only if $t = s$ and $x^t = y^s$. Then, we define the weight function as $\psi(t, x) = \exp\left(\|x\|_{C^\alpha([0,T];\mathbb{R}^n)}\right)$ such that $A_R := \{x \in C^\alpha([0, T]; \mathbb{R}^n);\ (t, x) \in K_R\}$ is by Arzelà-Ascoli in Theorem 1.16 compact with respect to the norm $\|\cdot\|_{C^0([0,T];\mathbb{R}^n)}$. In this case, $[0, T] \times A_R$ is by Tychonoff compact in the product topology of $[0, T] \times C^0([0, T]; \mathbb{R}^n)$ such that K_R is as quotient also compact with respect to $d_\infty : \Lambda_T^\alpha \times \Lambda_T^\alpha \to [0, \infty[$. Hence, (Λ_T^α, ψ) is a weighted space.

Example 3.2 (continued)**.**

(vi) For a measure space $(\Omega, \mathcal{F}, \mu)$ and some $1 \leqslant p < \infty$, let $X = L^p(\Omega)$ be the space of L^p-functions $x : \Omega \to \mathbb{R}$ such that the norm $\|x\|_{L^p(\Omega)} = \left(\int_\Omega |x(\omega)|^p d\mu(\omega)\right)^{1/p}$ is finite. Then, the dual of $L^q(\Omega)$ is isometrically isomorph to $L^p(\Omega)$, with $1/p + 1/q = 1$, such that $L^p(\Omega)$ can be equipped with a weak--topology. In this case, $\psi(x) = \exp\left(\|x\|_{L^p(\Omega)}\right)$ is by Banach-Alaoglu in Theorem 1.15 admissible and $(L^p(\Omega), \psi)$ is a weighted space. For $p = \infty$, the dual of $L^1(\Omega)$ is isometrically isomorph to $L^\infty(\Omega)$ if $\mu : \mathcal{F} \to [0, \infty[$ is σ-finite, where $L^\infty(\Omega) := \{x : \Omega \to \mathbb{R};\ x \text{ is } \mu\text{-a.e. bounded}\}$.*

(vii) For a bounded subset $\Omega \subset \mathbb{R}^n$ and the Lebesgue measure $\lambda : \mathcal{F} \to [0, \infty[$, we could consider a subset $X \subset L^p(\Omega) := L^p(\Omega, \mathcal{F}, \lambda)$ instead of $L^p(\Omega)$ consisting of functions $x : \Omega \to \mathbb{R}$ such that

$$\|x\|_p = \|x\|_{L^p(\Omega)} + \sup_{h \in \mathbb{R}^n \setminus \{0\}} \frac{\|x - x(\cdot + h)\|_{L^p(\mathbb{R}^n)}}{\|h\|}$$

is finite, where $x : \Omega \to \mathbb{R}$ is extended to $\mathbb{R}^n$ by setting it to zero outside of Ω. Similarly as in (iv), we equip X with the norm topology $\|\cdot\|_{L^p(\Omega)}$ instead of the weak--topology and define the weight function $\psi(x) = \exp(\|x\|_p)$. Then, $K_R = \psi^{-1}(]0, R])$ is by Fréchet-Kolmogorov in Corollary 1.18 compact with respect to $\|\cdot\|_{L^p(\Omega)}$, for all $R > 0$, such that (X, ψ) is a weighted space.*

From now on, $\psi : X \to]0, \infty[$ shall always denote an admissible weight function. Besides the vector space of bounded and continuous functions $C_b^0(X; Y)$, we introduce the vector space of Y-valued functions

$$B_\psi(X; Y) = \left\{ f : X \to Y;\ \sup_{x \in X} \frac{p(f(x))}{\psi(x)} < \infty,\ \text{for each } p \in \mathcal{P} \right\},$$

whose growth is controlled by the growth of $\psi : X \to]0, \infty[$, equipped with the topology τ_ψ^0 generated by the family of seminorms

$$\mathcal{Q} = \left\{ f \mapsto q(f) = \sup_{x \in X} p(f(x))/\psi(x);\ p \in \mathcal{P} \right\},$$

such that $(B_\psi(X; Y), \tau_\psi^0)$ is itself a locally convex topological vector space. Moreover, since τ_ψ^0 is a weaker topology on $C_b^0(X; Y)$ than the uniform topology τ_u^0 such that the embedding $C_b^0(X; Y) \subseteq B_\psi(X; Y)$ holds true.

Definition 3.3. *We define $\mathcal{B}_\psi(X;Y)$ as the τ_ψ^0-completion of $C_b^0(X;Y)$, which is a complete locally convex topological vector space with topology τ_ψ^0 generated by the family of seminorms $\mathcal{Q}$. If $Y=\mathbb{R}$, we write $\mathcal{B}_\psi(X)=\mathcal{B}_\psi(X;\mathbb{R})$, where the weighted norm $\|f\|_{\mathcal{B}_\psi(X)}=\sup_{x\in X}\frac{|f(x)|}{\psi(x)}$ turns $\mathcal{B}_\psi(X)$ into a Banach space.*

If X is compact, the weight function $\psi(x)=1$ is admissible. However, on general spaces, $\psi : X \to]0,\infty[$ grows on the compact preimages $K_R := \psi^{-1}(]0,R])$ such that elements of $\mathcal{B}_\psi(X;Y)$ are typically unbounded, but their growth is controlled by the growth of $\psi : X \to]0,\infty[$. For $Y=\mathbb{R}$, the following result characterizes functions in $\mathcal{B}_\psi(X)$, which relies on the Tietze extension theorem for the Stone-Čech compactification of the completely regular Hausdorff space X, where we follow the proof of [DT10, Theorem 2.7].

Lemma 3.4 ([DT10, Theorem 2.7])**.** *Let $f : X \to \mathbb{R}$ be a function. Then, $f\in\mathcal{B}_\psi(X)$ if and only if $f|_{K_R}\in C^0(K_R)$, for all $R>0$, and*

$$\lim_{R\to\infty}\sup_{x\in X\setminus K_R}\frac{|f(x)|}{\psi(x)}=0. \tag{3.2.1}$$

In particular, $f\in\mathcal{B}_\psi(X)$ for every $f\in C^0(X)$, where (3.2.1) *holds true.*

Proof. For sufficiency, let $f\in\mathcal{B}_\psi(X)$ and fix $\varepsilon>0$. By definition of $\mathcal{B}_\psi(X)$ there exists $g\in C_b^0(X)$ with $\|f-g\|_{\mathcal{B}_\psi(X)}<\varepsilon/2$ such that $|f(x)|/\psi(x)\leqslant \varepsilon/2+|g(x)|/\psi(x)$, for all $x\in X$. Choosing $R>2\varepsilon^{-1}\sup_{x\in X}|g(x)|$, we have

$$\sup_{x\in X\setminus K_R}\frac{|f(x)|}{\psi(x)}\leqslant\frac{\varepsilon}{2}+\sup_{x\in X\setminus K_R}\frac{|g(x)|}{\psi(x)}\leqslant\frac{\varepsilon}{2}+\frac{\sup_{x\in X}|g(x)|}{R}<\varepsilon,$$

showing that (3.2.1) holds true. In addition, it follows that $\sup_{x\in K_R}|f(x)-g(x)|\leqslant R\sup_{x\in X}|f(x)-g(x)|/\psi(x)\leqslant\varepsilon R/2$, for all $R>0$, such that $f|_{K_R}: K_R\to\mathbb{R}$ is continuous as the uniform limit of continuous functions.

For necessity, let $f : X\to\mathbb{R}$ such that $f|_{K_R}\in C^0(K_R)$, for all $R>0$, and (3.2.1). Define $f_n : X\to\mathbb{R}$ by $f_n(x)=\min(\max(f(x),-n),n)$, for $x\in X$ and $n\in\mathbb{N}$, such that $f|_{K_R}\in C^0(K_R)$ implies $f_n|_{K_R}\in C^0(K_R)$. By [Bou74, Proposition IX§1.3+4], (X,τ_X) can be embedded into its Stone-Čech compactification $(\breve{X},\tau_{\breve{X}})$, where K_R is also compact in $\breve{X}$. By the Tietze extension theorem in [Bou74, Théorème IX§4.2], there exists an extension $g_{n,R}\in C_b^0(X)$ with $g_{n,R}|_{K_R}=f_n|_{K_R}$ and $\sup_{x\in X}|g_{n,R}(x)|\leqslant n$. From this, it follows that

$$\|f_n-g_{n,R}\|_{\mathcal{B}_\psi(X)}\leqslant\sup_{x\in X\setminus K_R}\frac{|f_n(x)-g_{n,R}(x)|}{\psi(x)}\leqslant\frac{2n}{R}, \tag{3.2.2}$$

which implies that $f_n \in \mathcal{B}_\psi(X)$ because $g_{n,R} \in C_b^0(X) \subset \mathcal{B}_\psi(X)$ and the right-hand side of (3.2.2) can become arbitrarily small. Moreover, there exists by (3.2.1) some $R > 0$ with $\sup_{x \in X \setminus K_R} |f(x)|/\psi(x) < \varepsilon/2$, for which we choose some $n \in \mathbb{N}$ with $n > \sup_{x \in K_R} |f(x)|$ such that $f|_{K_R} = f_n|_{K_R}$. This implies

$$\|f - f_n\|_{\mathcal{B}_\psi(X)} \leqslant \sup_{x \in X \setminus K_R} \frac{|f(x) - f_n(x)|}{\psi(x)} \leqslant 2 \sup_{x \in X \setminus K_R} \frac{|f_n(x)|}{\psi(x)} < \varepsilon,$$

which shows that $f \in \mathcal{B}_\psi(X)$. □

In addition, there exists for every $f \in \mathcal{B}_\psi(X)$ with $\sup_{x \in X} f(x) > 0$ by [DT10, Theorem 2.8] some $x_0 \in X$ such that $f(x)/\psi(x) \leqslant f(x_0)/\psi(x_0)$, for all $x \in X$, showing the analogy with functions vanishing at infinity.

3.2.2 Weighted Stone-Weierstrass Theorem

In this section, we generalize the Stone-Weierstrass in Theorem 1.35 to our weighted setting with vector-valued functions, where X is a σ-compact completely regular Hausdorff space and the admissible weight function $\psi : X \to]0, \infty[$ controls the approximation outside of large compact sets.

However, we have to assume an additional condition related to Nachbin's definition of localisability (see [Nac65, Definition 4]), where the weighted approximation problem allows for more than one upper semicontinuous weight $v : X \to [0, \infty[$ (e.g. $v(x) = 1/\psi(x)$, for an admissible weight function $\psi : X \to]0, \infty[$) and can therefore represent other topologies.

Definition 3.5. *A subalgebra $\mathcal{A} \subset \mathcal{B}_\psi(X)$ is called* locally point separating (under $\psi : X \to]0, \infty[$) *if there exists a point separating vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ such that $x \mapsto \exp(|\widetilde{a}(x)|) \in \mathcal{B}_\psi(X)$, for all $\widetilde{a} \in \widetilde{\mathcal{A}}$.*

Remark 3.6. *A subalgebra $\mathcal{A} \subset \mathcal{B}_\psi(X)$ is locally point separating under any given admissible weight function $\psi : X \to]0, \infty[$ if the subalgebra $\mathcal{A} \subset C_b^0(X)$ is point separating and consists only of bounded functions. This is related to the so-called "bounded approximation problem", see [Nac65, Theorem 1].*

Theorem 3.7 (Stone-Weierstrass on $\mathcal{B}_\psi(X)$)**.** *For a subalgebra $\mathcal{A} \subset \mathcal{B}_\psi(X)$, the following are equivalent.*

(i) $\mathcal{A}$ is locally point separating and vanishes nowhere.

(ii) $\mathcal{A}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X)$.

Proof. In order to show that (ii) implies (i), we assume by contradiction that $\mathcal{A}$ is not locally point separating or does not vanish nowhere. If $\mathcal{A}$ is not locally point separating, there exists either a point separating vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ such that $x \mapsto \exp(|\widetilde{a}(x)|) \notin \mathcal{B}_\psi(X)$, for all $\widetilde{a} \in \widetilde{\mathcal{A}}$, or there exist distinct $x_1, x_2 \in X$ with $a(x_1) = a(x_2)$, for all $a \in \mathcal{A}$ satisfying $x \mapsto \exp(|a(x)|) \in \mathcal{B}_\psi(X)$. In the first case, we obtain a contradiction because $\widetilde{\mathcal{A}} \subset \mathcal{A}$ contains the zero function, whereas otherwise we define the proper closed subalgebra $\mathcal{A}_1 = \{f \in \mathcal{B}_\psi(X);\, f(x_1) = f(x_2)\}$ of $\mathcal{B}_\psi(X)$. In addition, if $\mathcal{A}$ does not vanish nowhere, i.e. there exists $x_0 \in X$ with $a(x_0) = 0$, for all $a \in \mathcal{A}$, we define the proper closed subalgebra $\mathcal{A}_2 = \{f \in \mathcal{B}_\psi(X);\, f(x_0) = 0\}$. Then, the τ_ψ^0-closure of $\mathcal{A}$ lies either in $\mathcal{A}_1$ or $\mathcal{A}_2$ such that the τ_ψ^0-closure of $\mathcal{A}$ is also a proper subalgebra of $\mathcal{B}_\psi(X)$, which contradicts the assertion in (ii) such that the assumptions in (i) must be satisfied.

For the other direction, it suffices to show the approximation of some $f \in C_b^0(X)$ by an element $a \in \mathcal{A}$, because $C_b^0(X)$ is by definition τ_ψ^0-dense in $\mathcal{B}_\psi(X)$. We first assume that $\mathcal{A}$ consists only of bounded functions in $C_b^0(X)$ such that the assumption of $\mathcal{A}$ being locally point separating reduces to $\mathcal{A}$ being point separating. Let $f \in C_b^0(X)$, fix $\varepsilon > 0$, and define $M = (\inf_{x\in X} \psi(x))^{-1} > 0$ as well as $b = \sup_{x\in X} |f(x)| + \varepsilon/(4M)$. Choosing $R \geqslant 4b/\varepsilon$, we observe that the restriction $\mathcal{A}|_{K_R}$ is a point separating subalgebra of $C^0(K_R)$, which vanishes nowhere, such that there exists by the classical Stone-Weierstrass in Theorem 1.35 some $a \in \mathcal{A}$ with

$$\sup_{x\in K_R} |f(x) - a(x)| < \frac{\varepsilon}{4M},$$

which implies $|a(x)| \leqslant \varepsilon/(4M) + \sup_{x\in X} |f(x)| = b$, for all $x \in K_R$. Let $g \in C_b^0(\mathbb{R})$ be the function defined by $g(s) = \max(\min(s, b), -b)$, for $s \in \mathbb{R}$, such that $g(a(x)) = a(x)$, for all $x \in K_R$. Then, it follows that

$$\begin{aligned} \|f - g \circ a\|_{\mathcal{B}_\psi(X)} &\leqslant M \sup_{x\in K_R} |f(x) - a(x)| + \sup_{x\in X\setminus K_R} \frac{|f(x)|}{\psi(x)} + \sup_{x\in X\setminus K_R} \frac{|g(a(x))|}{\psi(x)} \\ &< M\frac{\varepsilon}{4M} + \frac{\sup_{x\in X} |f(x)|}{R} + \frac{\sup_{x\in X} |g(a(x))|}{R} \\ &\leqslant \frac{\varepsilon}{4} + \frac{b}{R} + \frac{b}{R} \leqslant \frac{3\varepsilon}{4}. \end{aligned} \tag{3.2.3}$$

For $c = \sup_{x\in X} |a(x)|$, we find by Weierstrass' approximation result in Theorem 1.33 some polynomial $p_n \in \mathrm{Pol}(\mathbb{R})$ satisfying $|g(s) - p_n(s)| < \varepsilon/(4M)$, for

all $|s| \leqslant c$. Using this, we observe that

$$\begin{aligned}\|(g-p_n)\circ a\|_{\mathcal{B}_\psi(X)} &\leqslant M \sup_{x\in X} |g(a(x)) - p_n(a(x))| \leqslant M \sup_{|s|\leqslant c} |g(s)-p_n(s)| \\ &< M\frac{\varepsilon}{4M} = \frac{\varepsilon}{4}.\end{aligned} \tag{3.2.4}$$

Combining (3.2.3) with (3.2.4) and using that $p_n \circ a \in \mathcal{A}$ because $\mathcal{A}$ is a subalgebra, we conclude that

$$\|f - p_n\circ a\|_{\mathcal{B}_\psi(X)} \leqslant \|f - g\circ a\|_{\mathcal{B}_\psi(X)} + \|(g-p_n)\circ a\|_{\mathcal{B}_\psi(X)} \leqslant \frac{3\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.$$

Since $\varepsilon > 0$ and $f \in C_b^0(X)$ were chosen arbitrarily, it follows that $\mathcal{A} \subset C_b^0(X)$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X)$.

For the general case of a locally point separating subalgebra $\mathcal{A} \subset \mathcal{B}_\psi(X)$ with point separating vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ such that $x \mapsto \exp(|\widetilde{a}(x)|) \in \mathcal{B}_\psi(X)$, for all $\widetilde{a} \in \widetilde{\mathcal{A}}$, we first show that $x \mapsto \cos(\widetilde{a}(x))$ and $x \mapsto \sin(\widetilde{a}(x))$, with $\widetilde{a} \in \widetilde{\mathcal{A}}$, belong to the τ_ψ^0-closure of $\mathcal{A}$. Indeed, let $\widetilde{a} \in \widetilde{\mathcal{A}}$ and fix some $\varepsilon > 0$, for which there exists by Lemma 3.4 some $R \geqslant 4/\varepsilon$ such that

$$\sup_{x\in X\setminus K_R} \frac{\exp(|\widetilde{a}(x)|)}{\psi(x)} \leqslant \frac{\varepsilon}{4}.$$

Let $c = \sup_{x\in K_R} |\widetilde{a}(x)|$ such that the Taylor polynomial of the cosinus $p_n(s) = \sum_{k=0}^n \frac{(-1)^k}{(2k)!} s^{2k}$ satisfies $\sup_{|s|\leqslant c} |\cos(s) - p_n(s)| \leqslant \frac{c^{2n+1}}{(2n+1)!}$. Choosing $n \in \mathbb{N}$ large enough such that $\frac{c^{2n+1}}{(2n+1)!} \leqslant \frac{\varepsilon}{2M}$ and using that $|p_n(s)| \leqslant \exp(|s|)$, we conclude that

$$\begin{aligned}\|(\cos - p_n)\circ \widetilde{a}\|_{\mathcal{B}_\psi(X)} &\leqslant M \sup_{x\in K_R} |\cos(\widetilde{a}(x)) - p_n(\widetilde{a}(x))| \\ &\quad + \sup_{x\in X\setminus K_R} \frac{|\cos(\widetilde{a}(x))|}{\psi(x)} + \sup_{x\in X\setminus K_R} \frac{|p_n(\widetilde{a}(x))|}{\psi(x)} \\ &< M\frac{\varepsilon}{2M} + \frac{\sup_{x\in X}|\cos(\widetilde{a}(x))|}{R} + \sup_{x\in X\setminus K_R} \frac{\exp(|\widetilde{a}(x)|)}{\psi(x)} \\ &\leqslant \frac{\varepsilon}{2} + \frac{1}{R} + \frac{\varepsilon}{4} \leqslant \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ was chosen arbitrarily, it follows that $x \mapsto \cos(\widetilde{a}(x))$ belongs to the τ_ψ^0-closure of $\mathcal{A}$, which holds analogously true for the function $x \mapsto \sin(\widetilde{a}(x))$,

for all $\widetilde{a} \in \widetilde{\mathcal{A}}$. Hence, we conclude that the subalgebra

$$\mathcal{A}_{\mathrm{trig}} = \left\{ x \mapsto \lambda_1 \cos\left(\widetilde{a}_1(x)\right) + \lambda_2 \sin\left(\widetilde{a}_2(x)\right) ;\, \lambda_1, \lambda_2 \in \mathbb{R},\, \widetilde{a}_1, \widetilde{a}_2 \in \widetilde{\mathcal{A}} \right\}$$

of $C_b^0(X) \subset \mathcal{B}_\psi(X)$ is point separating, vanishes nowhere and lies in the τ_ψ^0-closure of $\mathcal{A}$. By applying the above arguments to the subalgebra $\mathcal{A}_{\mathrm{trig}} \subset C_b^0(X)$ consisting of bounded functions, we conclude that $\mathcal{A}_{\mathrm{trig}}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X)$. However, because $\mathcal{A}_{\mathrm{trig}}$ lies in the τ_ψ^0-closure of $\mathcal{A}$, $\mathcal{A}$ is also τ_ψ^0-dense in $\mathcal{B}_\psi(X)$. □

For the vector-valued case, we state the vector-valued Stone-Weierstrass of R. Creighton Buck in [Buc58] with compact Hausdorff domain, for which we introduce the notion of a submodule.

Definition 3.8. *Let $\mathcal{A} \subset C^0(X)$ be a subalgebra and let $\mathcal{W} \subset C^0(X;Y)$ be a vector subspace. Then, $\mathcal{W}$ is called an* $\mathcal{A}$-submodule *if $a \cdot w \in \mathcal{W}$, for all $a \in \mathcal{A}$ and $w \in \mathcal{W}$, where $x \mapsto (a \cdot w)(x) := a(x)w(x)$.*

Theorem 3.9 (Stone-Weierstrass, [Buc58, p. 103])**.** *Let (X, τ_X) be a compact Hausdorff space and let (Y, τ_Y) be a locally convex topological vector space. Assume that $\mathcal{A} \subset C^0(X)$ is a subalgebra and let $\mathcal{W} \subset C^0(X;Y)$ be an $\mathcal{A}$-submodule. Then, the following are equivalent.*

(i) $\mathcal{A}$ is point separating and vanishes nowhere, and $\mathcal{W}(x) := \{w(x);\, w \in \mathcal{W}\}$ is dense in Y, for all $x \in X$.

(ii) $\mathcal{W}$ is dense in $C^0(X;Y)$.

For the vector-valued Stone-Weierstrass on weighted spaces, we consider the weight functions $\psi_{w,p} : X \to]0,\infty[$ defined by $\psi_{w,p}(x) = \psi(x)/(1 + p(w(x)))$, for $x \in X$ and some $w \in \mathcal{B}_\psi(X;Y)$ and $p \in \mathcal{P}$.

Lemma 3.10. *For every $w \in \mathcal{B}_\psi(X;Y)$ and $p \in \mathcal{P}$, the weight function $\psi_{w,p} : X \to]0,\infty[$ is admissible.*

Proof. For $w \in \mathcal{B}_\psi(X;Y)$ and $p \in \mathcal{P}$, it follows that $x \mapsto 1 + p(w(x)) \in \mathcal{B}_\psi(X)$ such that Lemma 3.4 implies $\lim_{R\to\infty} \sup_{x \in X \setminus K_R} (1 + p(w(x)))/\psi(x) = 0$. Using this, we find for every $R > 0$ some $\widetilde{R} \geqslant R$ such that $x \in K_{w,p,R} := \psi_{w,p}^{-1}(]0,R])$, i.e. $\psi_{w,p}(x) \leqslant R$, implies $x \in K_{\widetilde{R}} = \psi^{-1}(]0,\widetilde{R}])$. Hence, $K_{w,p,R} \subseteq K_{\widetilde{R}}$ is by continuity of $x \mapsto 1 + p(w(x))$ closed and thus compact, which shows that $\psi_{w,p} : X \to]0,\infty[$ is admissible. □

Theorem 3.11 (Stone-Weierstrass on $\mathcal{B}_\psi(X;Y)$)**.** *Let $\mathcal{A} \subset \mathcal{B}_\psi(X)$ be a subalgebra and assume that $\mathcal{W} \subset \mathcal{B}_\psi(X;Y)$ is an $\mathcal{A}$-submodule. Then, the following are equivalent.*

(i) $\mathcal{A}$ vanishes nowhere and is locally point separating under every weight function $\psi_{w,p} : X \to]0,\infty[$, for all $w \in \mathcal{W}$ and $p \in \mathcal{P}$. In addition, $\mathcal{W}(x) := \{w(x);\, w \in \mathcal{W}\}$ is dense in Y, for all $x \in X$.

(ii) $\mathcal{W}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$.

Proof. In order to show that (ii) implies (i), we assume by contradiction that $\mathcal{A}$ is not locally point separating under some $\psi_{w,p} : X \to]0,\infty[$, does not vanish nowhere, or there exists some $x_3 \in X$ such that $\mathcal{W}(x_3) \subset E$, for a proper closed vector subspace $E \subset Y$. If $\mathcal{A}$ is not locally point separating under $\psi_{w,p} : X \to]0,\infty[$, there exists either a point separating vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ with $x \mapsto \exp(|\widetilde{a}(x)|) \notin \mathcal{B}_{\psi_{w,p}}(X)$, for some $\widetilde{a} \in \widetilde{\mathcal{A}}$, or there exist distinct $x_1, x_2 \in X$ with $a(x_1) = a(x_2)$, for all $a \in \mathcal{A}$ satisfying $x \mapsto \exp(|a(x)|) \in \mathcal{B}_{\psi_{w,p}}(X)$. In the first case, we obtain a contradiction because $\widetilde{\mathcal{A}} \subset \mathcal{A}$ is a vector subspace such that $\widetilde{\mathcal{A}}$ contains the zero function, whereas in the second case we define the proper closed $\mathcal{A}$-submodule $\mathcal{W}_1 = \operatorname{span}\{a \cdot f;\, f \in \mathcal{B}_\psi(X;Y),\, a \in \mathcal{A},\, a(x_1) = a(x_2)\}$. Moreover, if $\mathcal{A}$ does not vanish nowhere, i.e. there exists some $x_0 \in X$ with $a(x_0) = 0$, for all $a \in \mathcal{A}$, or $\mathcal{W}(x_3)$ is containted in E, we define the proper closed $\mathcal{A}$-submodules $\mathcal{W}_2 = \operatorname{span}\{a \cdot f;\, f \in \mathcal{B}_\psi(X;Y),\, a \in \mathcal{A},\, a(x_0) = 0\}$ and $\mathcal{W}_3 = \operatorname{span}\{a \cdot f;\, f \in \mathcal{B}_\psi(X;Y),\, a \in \mathcal{A},\, f(x_3) \in E\}$. Then, the τ_ψ^0-closure of $\mathcal{W}$ lies either in $\mathcal{W}_1$, $\mathcal{W}_2$ or $\mathcal{W}_3$, such that the τ_ψ^0-closure of $\mathcal{W}$ is a proper closed $\mathcal{A}$-submodule of $\mathcal{B}_\psi(X;Y)$, which contradicts (ii) such that the assumptions (i) must be satisfied.

For the other direction, we first show that the τ_ψ^0-closure of $\mathcal{W}$ is a $C_b^0(X)$-submodule. Let $g \in C_b^0(X)$ and let $\bar{w} : X \to Y$ belong to the τ_ψ^0-closure of $\mathcal{W}$, fix some $\varepsilon > 0$ and $p \in \mathcal{P}$, and define $c = \sup_{x\in X} |g(x)|$. Then, there exists some $w \in \mathcal{W}$ with $q(\bar{w} - w) = \sup_{x\in X} p(\bar{w}(x) - w(x))/\psi(x) < \varepsilon/(2c)$, which implies

$$q(g \cdot \bar{w} - g \cdot w) = \sup_{x\in X} \frac{|g(x)| p(\bar{w}(x) - w(x))}{\psi(x)} \leqslant c \sup_{x\in X} \frac{p(\bar{w}(x) - w(x))}{\psi(x)} < \frac{\varepsilon}{2}, \tag{3.2.5}$$

where $q \in \mathcal{Q}$ denotes the corresponding seminorm. Moreover, $\psi_{w,p} : X \to]0,\infty[$ is by Lemma 3.10 admissible and for every $a \in \mathcal{A}$, we have $a \cdot w \in \mathcal{B}_\psi(X;Y)$ such that $x \mapsto |a(x)|(1 + p(w(x))) \in \mathcal{B}_\psi(X)$. Then, it follows by Lemma 3.4 that $a \in \mathcal{B}_{\psi_{w,p}}(X)$ such that $\mathcal{A}$ is also a subalgebra of $\mathcal{B}_{\psi_{w,p}}(X)$. Hence, by applying the Stone-Weierstrass in Theorem 3.7 on $\mathcal{B}_{\psi_{w,p}}(X)$ there exists some

$a \in \mathcal{A}$ with $\|g - a\|_{\mathcal{B}_{\psi_{w,p}}(X)} < \varepsilon/2$ such that

$$\begin{aligned} q\left(g \cdot w - a \cdot w\right) &= \sup_{x \in X} \frac{p(g(x)w(x) - a(x)w(x))}{\psi(x)} \\ &\leqslant \sup_{x \in X} \frac{|g(x) - a(x)|(1 + p(w(x)))}{\psi(x)} \\ &= \|g - a\|_{\mathcal{B}_{\psi_{w,p}}(X)} < \frac{\varepsilon}{2}. \end{aligned} \tag{3.2.6}$$

By combining (3.2.5) with (3.2.6), we conclude that $q\left(g \cdot \bar{w} - a \cdot w\right) < \varepsilon$, which shows that $g \cdot \bar{w} : X \to Y$ belongs to the τ_ψ^0-closure of $\mathcal{W}$, because $\varepsilon > 0$ and $p \in \mathcal{P}$ were chosen arbitrarily, and $\mathcal{W}$ is an $\mathcal{A}$-submodule. Since $g \in C_b^0(X)$ and $w \in \mathcal{W}$ were also chosen arbitrarily, the τ_ψ^0-closure of $\mathcal{W}$ is a $C_b^0(X)$-submodule.

Finally, in order to show (ii) we first observe that $C_b^0(X;Y)$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$ such that we only need to approximate some $f \in C_b^0(X;Y)$ by an element $w \in \mathcal{W}$. Let $f \in C_b^0(X;Y)$, fix $\varepsilon > 0$ and $p \in \mathcal{P}$, and define $b = \sup_{x \in X} p(f(x))$ as well as $M = \left(\inf_{x \in X} \psi(x)\right)^{-1}$. For $R \geqslant \max(4(b+1)/\varepsilon, 1/M)$, we observe that the restricted subalgebra $\mathcal{A}|_{K_R}$ of $C^0(K_R)$ and the restricted $\mathcal{A}|_{K_R}$-submodule $\mathcal{W}|_{K_R}$ of $C^0(K_R;Y)$ satisfy the assumptions of the classical Stone-Weierstrass in Theorem 3.9 such that there exists some $w \in \mathcal{W}$ with

$$\sup_{x \in K_R} p\left(f(x) - w(x)\right) < \frac{\varepsilon}{4M}. \tag{3.2.7}$$

Define $c = \sup_{x \in K_R} p(w(x))$, for which it follows by (3.2.7) that $c \leqslant \varepsilon/(4M) + b$. From this, we define the function $g \in C_b^0(X)$ by $g(x) = \min(1 + c, 1 + p(w(x)))/(1 + p(w(x)))$, for $x \in X$, such that $g(x) = 1$, for all $x \in K_R$. Then, the function $g \cdot w \in C_b^0(X;Y)$ is bounded with $\sup_{x \in X} p(g(x)w(x)) \leqslant 1 + c \leqslant 1 + b + \varepsilon/(4M)$ and belongs to the τ_ψ^0-closure of $\mathcal{W}$, because the latter is a $C_b^0(X)$-submodule. Using (3.2.7), it follows that

$$\begin{aligned} q(f - g \cdot w) &\leqslant M \sup_{x \in K_R} p(f(x) - w(x)) \\ &\quad + \sup_{x \in X \setminus K_R} \frac{p(f(x))}{\psi(x)} + \sup_{x \in X \setminus K_R} \frac{p(g(x)w(x))}{\psi(x)} \\ &< M \frac{\varepsilon}{4M} + \frac{\sup_{x \in X} p(f(x))}{R} + \frac{\sup_{x \in X} p(g(x)w(x))}{R} \\ &\leqslant \frac{\varepsilon}{4} + \frac{b}{R} + \frac{1+b}{R} + \frac{\varepsilon}{4MR} \leqslant \varepsilon, \end{aligned}$$

where $q \in \mathcal{Q}$ is the seminorm corresponding to p. Since $\varepsilon > 0$, $p \in \mathcal{P}$ and $f \in C_b^0(X;Y)$ were chosen arbitrarily, $\mathcal{W}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$. □

Remark 3.12. *In Theorem 3.11, $\mathcal{A} \subset \mathcal{B}_\psi(X)$ is locally point separating under every $\psi_{w,p} : X \to]0,\infty[$, for all $w \in \mathcal{W}$ and $p \in \mathcal{P}$, if either $\mathcal{A} \subset C_b^0(X)$ consists only of bounded functions (see Remark 3.6) or if $\mathcal{A} \subset \mathcal{B}_{\sqrt{\psi}}(X)$ is locally point separating under $\sqrt{\psi} : X \to]0,\infty[$ and $\mathcal{W}$ is a subset of $\mathcal{B}_{\sqrt{\psi}}(X;Y)$. In both cases, the multiplication $\mathcal{A} \times \mathcal{W} \ni (a,w) \mapsto a \cdot w \in \mathcal{B}_\psi(X;Y)$ becomes jointly continuous.*

While the real-valued Stone-Weierstrass in Theorem 3.7 can be applied to the weighted polynomial approximation problem of Sergei Bernstein in [Ber24], the vector-valued Stone-Weierstrass in Theorem 3.11 is a special case of [Nac65, Theorem 6], which has been generalized by João B. Prolla in [Pro71] extending the measure theoretic version of Erret Bishop in [Bis61] for the compact case. In addition, Silvio Machado derived in [Mac77] a quantitative version relying on Zorn's lemma, see also [BD81] and [Ran84].

3.3 Neural Networks in Infinite Dimension

In [MP43], [Cyb89] and [Hor91], neural networks were introduced between finite dimensional Euclidean spaces, whereas in the following, we assume that the input space (X,ψ) is given by a weighted space, the hidden layer space (Z,ψ_Z) is a complete weighted topological vector space, whose topology τ_Z is locally convex, and the output space (Y,τ_Y) is given by a complete locally convex topological vector space.

3.3.1 Abstract Neural Networks

First, we introduce the infinite dimensional analogue of the weight matrices and bias vectors used in neural networks between finite dimensional spaces, followed by the definition of the abstract neural network.

Definition 3.13. *A subset $\mathcal{H} \subset \mathcal{B}_\psi(X;Z)$ is called an* additive family (between X and Z) *if $\mathcal{H}$ is closed under addition, point separating and contains the constant functions.*

Remark 3.14. *In the finite dimensional case of $X = \mathbb{R}^n$ and $Z = \mathbb{R}^m$, an additive family is given by $\mathcal{H} = \{x \mapsto Ax + b;\, A \in \mathbb{R}^{m\times n},\, b \in \mathbb{R}^m\}$, showing the analogy to affine maps with weight matrices and bias vectors, whereas $\mathcal{H} = \{x \mapsto Ax + b;\, A \in \mathbb{N}^{m\times n},\, b \in \mathbb{R}^m\}$ is an even smaller additive family.*

Definition 3.15. *For an additive family $\mathcal{H} \subset \mathcal{B}_\psi(X;Z)$, a subset $\mathcal{R} \subset \mathcal{B}_{\psi_Z}(Z;Z)$ called the* activation family*, and a subset $\mathcal{L} \subset L(Z;Y)$ called the* linear readout*, we define an* abstract neural network (ANN) *$\varphi : X \to Y$ by*

$$\varphi(x) = \sum_{k=1}^{N} \beta_k \rho_k(h_k(x)), \tag{3.3.1}$$

for $x \in X$, where $N \in \mathbb{N}_{\geqslant 1}$, $\beta_1, ..., \beta_N \in \mathcal{L}$ denote the linear readouts*, $\rho_1, ..., \rho_N \in \mathcal{R}$ are the* activation functions *and $h_1, ..., h_N \in \mathcal{H}$ represent the* hidden layer maps*. Moreover, $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ denotes the set of ANNs of the form* (3.3.1)*, where the triple $(\mathcal{H}, \mathcal{R}, \mathcal{L})$ is its* architecture (between X and Y)*, see Figure 3.1.*

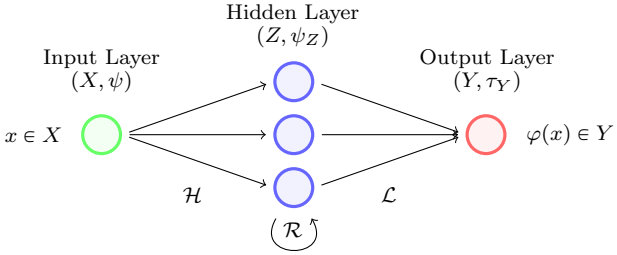


Figure 3.1: An example of an abstract neural network $\varphi : X \to Y$ with architecture $(\mathcal{H}, \mathcal{R}, \mathcal{L})$ and $N = 3$ number of nodes.

Remark 3.16. *In [Cyb89] and [Hor91], neural networks were introduced between $X = \mathbb{R}^n$ and $Y = \mathbb{R}$ with $Z = \mathbb{R}$, where the additive family $\mathcal{H}$ consists of affine maps $h(x) = a^\top x + b$, with $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the activation family $\mathcal{R} = \{\rho\}$ is a singleton $\rho : \mathbb{R} \to \mathbb{R}$, and the linear readout $\mathcal{L}$ contains functions $\beta(z) = wz$, with $w \in \mathbb{R}$. Then, $\mathcal{NN}_{\mathbb{R}^n,\mathbb{R}}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is dense in $C^0(\mathbb{R}^n)$ on compacta if $\rho : \mathbb{R} \to \mathbb{R}$ is continuous and either*

(i) bounded and sigmoidal (or discriminatory in the sense of Definition 2.2, see [Cyb89, Theorem 1]),

(ii) bounded and non-constant (see [Hor91, Theorem 2]),

Remark 3.16 (continued)**.**

(iii) non-polynomial (see [Pin99, Theorem 3.1]), or

(iv) non-affine and differentiable at one point with non-zero derivative (see [KL20, Theorem 3.2]).

If $\rho : \mathbb{R} \to \mathbb{R}$ is additionally bounded, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}}$ is τ^0_u-dense in $C^0_b(\mathbb{R}^n)$ because τ^0_u is a weaker topology on $C^0_b(\mathbb{R}^n)$ than the topology of uniform convergence on compacts. Finally, by using that τ^0_ψ is again a weaker topology on $C^0_b(\mathbb{R}^n)$, which is τ^0_ψ-dense in $\mathcal{B}_\psi(\mathbb{R}^n)$, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}}$ is also τ^0_ψ-dense in $\mathcal{B}_\psi(\mathbb{R}^n)$.

For the infinite dimensional generalization of neural networks described in Remark 3.16, we first consider the case with a multidimensional output $Y = \mathbb{R}^d$ followed by some examples with infinite dimensional spaces.

Example 3.17 (Learning a Function)**.** *An abstract neural network $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ with $Z = \mathbb{R}^d$ can be obtained by stacking d real-valued neural networks on top of each other such that $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ is given by*

$$\varphi(x) = \begin{pmatrix} \sum_{k=1}^N \beta_k^1 \rho_0\left(\left(a_k^1\right)^\top x + b_k^1\right) \\ \vdots \\ \sum_{k=1}^N \beta_k^d \rho_0\left(\left(a_k^d\right)^\top x + b_k^d\right) \end{pmatrix} = \sum_{k=1}^N \beta_k \rho\left(A_k x + b_k\right),$$

where $\mathcal{L}$ consists of diagonal matrices $\beta_k = diag(\beta_k^1, ..., \beta_k^d)$ and $\mathcal{H}$ contains hidden layer maps $h(x) = A_k x + b_k$ with $A_k = (a_k^1, ..., a_k^d)^\top \in \mathbb{R}^{d\times n}$ and $b_k \in \mathbb{R}^d$. In addition, the activation function $\rho : \mathbb{R}^d \to \mathbb{R}^d$ is obtained by componentwise application of some $\rho_0 : \mathbb{R} \to \mathbb{R}$ used in Remark 3.16, i.e. $\rho(z) = (\rho_0(z_1), ..., \rho_0(z_d))^\top$.

On the other hand, a fully connected neural network $\varphi : \mathbb{R}^n \to \mathbb{R}^d$ with $Z = \mathbb{R}^m$ is obtained with $\mathcal{H} = \{x \mapsto Ax + b;\, A \in \mathbb{R}^{m\times n},\, b \in \mathbb{R}^m\}$ and $\mathcal{L} = \{z \mapsto Wz;\, W \in \mathbb{R}^{d\times m}\}$.

Example 3.18 (Learning a Functional)**.** *For a weighted space (X, ψ) and $Y = Z = \mathbb{R}$, an abstract neural network $\varphi : X \to \mathbb{R}$ consists of hidden layer maps $h : X \to \mathbb{R}$, a bounded and continuous activation function $\rho : \mathbb{R} \to \mathbb{R}$ described in Remark 3.16, and the linear readout is given by $\mathcal{L} = \{z \mapsto wz;\, w \in \mathbb{R}\}$.*

Example 3.19 (Learning an Operator)**.** *For a weighted space (X, ψ) and a complete locally convex topological vector space (Y, τ_Y), we choose $Z = \mathbb{R}^m$ such that the additive family of an abstract neural network $\varphi : X \to Y$ consists of hidden layer maps $h : X \to \mathbb{R}^m$ and the linear readouts are of the form $\beta(z) = y\, w^\top z$, for $w \in \mathbb{R}^m$ and $y \in Y$. Hereby, the activation function $\rho : \mathbb{R}^m \to \mathbb{R}^m$ is chosen as in Example 3.17.*

On the other hand, deep abstract neural networks with finitely many hidden layer can be constructed by concatenation as in the case with neural networks between Euclidean spaces, see e.g. [GBC16].

Remark 3.20. *For weighted locally convex topological vector spaces (Z_1, ψ_1) and (Z_2, ψ_2), let $\mathcal{H}$ be an additive family between X and Z_1, assume that $\mathcal{R}_1 \subset \mathcal{B}_{\psi_{Z_1}}(Z_1; Z_1)$ and $\mathcal{R}_2 \subset \mathcal{B}_{\psi_{Z_2}}(Z_2; Z_2)$, and suppose that $\mathcal{L}_1 \subset L(Z_1; Z_2)$ and $\mathcal{L}_2 \subset L(Z_2; Y)$ are linear readouts. If $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X; Z_2)$, it is an additive family between X and Z_2 such that an abstract neural network $\varphi : X \to Y$ with two hidden layers is of the form*

$$\varphi(x) = \sum_{k=1}^{N_2} \beta_k \rho_k \left(\varphi_k(x)\right) = \sum_{k=1}^{N_2} \beta_k \rho_k \left(\sum_{l=1}^{N_1} \beta_{kl} \rho_{kl}(h_{kl}(x)) \right),$$

for $x \in X$, where $\beta_k \in \mathcal{L}_2$, $\rho_k \in \mathcal{R}_2$ and $\varphi_k \in \mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$, with $\beta_{kl} \in \mathcal{L}_1$, $\rho_{kl} \in \mathcal{R}_1$, $h_{kl} \in \mathcal{H}$, see Figure 3.2.

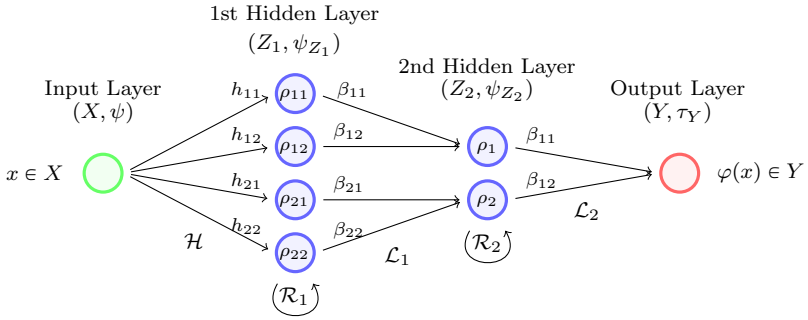


Figure 3.2: An example of an abstract neural network $\varphi : X \to Y$ with two hidden layers, where $(N_1, N_2) = (2, 2)$.

Analogously to Remark 3.20, it is possible to construct abstract neural networks with finitely many hidden layers, for which the results of the upcoming sections can be generalized by iterating the argument.

3.3.2 More Examples of Abstract Neural Networks

For a given input space (X, ψ), hidden layer space (Z, ψ_Z), and an output space (Y, τ_Y), finding an additive family $\mathcal{H}$ between X and Z is not always straightforward. If X is additionally a locally convex topological vector space, the additive family can be found with the help of continuous linear maps.

Lemma 3.21. *Let (X, ψ) and (Z, ψ_Z) be weighted topological vector spaces, where the topology τ_X of X is also locally convex. Then, $\mathcal{H} = \{x \mapsto \gamma(x)+b;\ \gamma \in L(X; Z), b \in Z\}$ is an additive family between X and Z.*

Proof. By definition, $\mathcal{H}$ is closed under addition and contains the constant functions. For point separating, let $x_1, x_2 \in X$ be distinct and define $E = \operatorname{span}\{x_1, x_2\}$. If x_1 and x_2 are linearly independent, we define the continuous linear functional $\widetilde{l} : E \to \mathbb{R}$ by $\widetilde{l}(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1$, for $\lambda_i \in \mathbb{R}$, such that $\widetilde{l}(x_1) = 1 \neq 0 = \widetilde{l}(x_2)$. Otherwise, if x_1 and x_2 are linearly dependent, we can assume without loss of generality that $x_1 \neq 0$ and $x_2 = \lambda x_1$, for some $\lambda \in \mathbb{R}\backslash\{0, 1\}$, such that the continuous linear functional $\widetilde{l} : E \to \mathbb{R}$ defined by $\widetilde{l}(\lambda x_1) = \lambda$, for $\lambda \in \mathbb{R}$, satisfies $\widetilde{l}(x_1) = 1 \neq \lambda = \widetilde{l}(x_2)$. In both cases, $\widetilde{l} : E \to \mathbb{R}$ can be extended by Hahn-Banach in Theorem 1.28 to a continuous linear functional $l : X \to \mathbb{R}$ such that $x \mapsto h(x) := l(x)z \in \mathcal{H}$ with $z \in Z\backslash\{0\}$ satisfies

$$h(x_1) = l(x_1)z = \widetilde{l}(x_1)z \neq \widetilde{l}(x_2)z = l(x_2)z = h(x_2).$$

Since $x_1, x_2 \in X$ were chosen arbitrarily, it follows that $\mathcal{H}$ is point separating on X. □

Lemma 3.21 shows in particular that an abstract neural network over a locally convex topological vector space (X, τ_X) with hidden and output layer $Z = Y = \mathbb{R}$ is given by some function $\varphi : X \to \mathbb{R}$ of the form

$$\varphi(x) = \sum_{k=1}^{N} \beta_k \rho\left(l_k(x) + b_k\right),$$

for $x \in X$, where $l_k \in X^*$, $\beta_k, b_k \in \mathbb{R}$, and $\rho : \mathbb{R} \to \mathbb{R}$ is an activation function as in Remark 3.16. In the following, we revisit the introductory examples of Section 3.1 and consider some additive families.

Example 3.22. *We consider the abstract neural networks of the motivational examples in Section 3.1 again.*

(i) For a compact subset $\Omega \subset \mathbb{R}^n$, let $X = C^0(\Omega)$ and $Z = Y = \mathbb{R}$. Then, an additive family is given by

$$\mathcal{H} = \left\{ C^0(\Omega) \ni x \mapsto \int_\Omega x(u) d\nu(u) + b \in \mathbb{R};\ \nu : \mathcal{F} \to \mathbb{R},\ b \in \mathbb{R} \right\},$$

with $\nu : \mathcal{F} \to \mathbb{R}$ denoting a finite signed Radon measure defined on the Borel σ-algebra of Ω, whereas the activation family $\mathcal{R}$ consists of activation functions $\rho : \mathbb{R} \to \mathbb{R}$ described in Remark 3.16.

(ii) Let $X = \Lambda_T$ be the space of stopped paths and let $Z = Y = \mathbb{R}$. Then, an additive family is given by

$$\mathcal{H} = \left\{ \Lambda_T \ni (t,x) \mapsto at + \sum_{i=1}^n \int_0^t x_i(u) d\nu_i(u) + b \in \mathbb{R};\ \begin{matrix} \nu_i : \mathcal{F} \to \mathbb{R} \\ a, b \in \mathbb{R} \end{matrix} \right\},$$

with $\nu_i : \mathcal{F} \to \mathbb{R}$ denoting a finite signed Radon measure defined on the Borel σ-algebra of $[0,T]$, whereas the activation family $\mathcal{R}$ consists of activation functions $\rho : \mathbb{R} \to \mathbb{R}$ described in Remark 3.16.

(iii) Using the Sobolev space $W^{1,p}(\Omega)$ such that $X = Y = Z = W^{1,p}(\mathbb{R}^n)$, an additive family is given by

$$\mathcal{H} = \left\{ W^{1,p}(\mathbb{R}^n) \ni x_0 \mapsto a * x_0 + b;\ a \in C_c^\infty(\mathbb{R}^n),\ b \in W^{1,p}(\mathbb{R}^n) \right\},$$

where the convolution $a * x_0 \in W^{1,p}(\mathbb{R}^n)$ is by [Bré11, Lemma 9.1] well-defined. In this case, an activation function $\rho : W^{1,p}(\mathbb{R}^n) \to W^{1,p}(\mathbb{R}^n)$ can be obtained by the pointwise evaluation of a sigmoidal function $\rho_0 \in C_b^1(\mathbb{R})$ with bounded derivatives and $\rho_0(0) = 0$, where [Bré11, Proposition 9.5] shows that $\rho(z) := (u \mapsto \rho_0(z(u))) \in W^{1,p}(\mathbb{R}^n)$, for all $z \in W^{1,p}(\mathbb{R}^n)$. Similar results can be obtained for $W^{k,p}(\mathbb{R}^n)$ with $kp \geqslant n$ by using the Sobolev or Nirenberg inequality, see [Ada76, p. 127].

Proof. For (i), the space of Radon measures is by Riesz representation theorem isomorphic to $X^* = L(X, \mathbb{R})$ such that the conclusion follows from Lemma 3.21. For (ii), we cannot apply Lemma 3.21 because Λ_T is not a vector space. However, $\mathcal{H}$ is closed under addition, contains the constant functions and for distinct $(t_1, x_1), (t_2, x_2) \in \Lambda_T$, we observe that if $t_1 \neq t_2$, the map $(t,x) \mapsto t \in \mathcal{H}$ separates (t_1, x_1) from (t_2, x_2). Otherwise, if $t_1 = t_2$ and $x_1(u_0) \neq x_2(u_0)$, for some $u_0 \in [0,T]$, it follows by continuity that there exists a Radon measure ν with

mass around u_0 such that the map $(t,x) \mapsto \sum_{i=1}^n \int_0^t x_i(u)d\nu_i(u) \in \mathcal{H}$ separates (t_1,x_1) from (t_2,x_2). For (iii), $\mathcal{H}$ is closed under addition, contains the constant functions and for distinct $x_1, x_2 \in W^{1,p}(\mathbb{R}^n)$, there exists a measurable set $E \subset \mathbb{R}^n$ with positive measure such that $x_1(u) > x_2(u)$ or $x_1(u) < x_2(u)$, for all $u \in E$. Then, by using $x \mapsto h(x) = \mathbb{1}_E * x \in \mathcal{H}$, it follows that

$$\begin{aligned} h(x_1) - h(x_2) = \mathbb{1}_E * (x_1 - x_2) &= \left(v \mapsto \int_E (x_1(v-u) - x_2(v-u))du \right) \\ &= \left(v \mapsto -\int_{v-E} (x_1(u) - x_2(u))du \right) \end{aligned}$$

is a non-zero function in $W^{1,p}(\mathbb{R}^n)$, because the right-hand side is non-zero, for all $v \in \mathbb{R}^n$ with $v - E \subseteq E$. □

On the other hand, we can also consider the dual of a Hausdorff locally convex topological vector space (S, τ_S) as input space (X, τ_X), where the topology τ_X of X is defined as the weak-$*$-topology.

Lemma 3.23. *Let (S,τ_S) be a non-trivial locally convex topological vector spaces and define (X,τ_X) as the dual of (S,τ_S) equipped with the weak-$*$-topology. Then, the set $\mathcal{H} = \{x \mapsto \alpha(x) + b;\ \alpha \in L(X;Z),\ b \in Z\}$ is an additive family between X and Z.*

Proof. By definition, $\mathcal{H}$ is closed under addition and contains the constant functions, whereas for point separating, we first observe that $S \neq \{0\}$ implies $X \neq \{0\}$. Indeed, choose some $s \in S\backslash\{0\}$, let $E = \operatorname{span}\{s\}$ and define $l \in E^*$ by $l(s) = 1$. Then, $l : E \to \mathbb{R}$ can be extended by Hahn-Banach in Theorem 1.28 to some non-zero $x \in X = S^*$, showing that $X \neq \{0\}$. Now, let $x_1, x_2 \in X$ be distinct points such that there exists some $s \in S$ with $x_1(s) \neq x_2(s)$. Then, we conclude for $x \mapsto h(x) := (s \mapsto x(s)z) \in \mathcal{H}$ with some $z \in Z\backslash\{0\}$ that

$$h(x_1)(s) = x_1(s)z \neq x_2(s)z = h(x_2)(s),$$

showing that $\mathcal{H}$ is point separating on X, because $x_1, x_2 \in X$ were arbitrary. □

Lemma 3.23 shows that an abstract neural network $\varphi : S^* \to \mathbb{R}$ over the dual S^* of a locally convex topological vector space (S,τ_S) equipped with the weak-$*$-topology and with hidden and output space $Z = Y = \mathbb{R}$ is given by

$$\varphi(x) = \sum_{k=1}^N \beta_k \rho\left(x(s_k) + b_k\right),$$

for $x \in S^*$, where $s_k \in S$, $\beta_k, b_k \in \mathbb{R}$, and $\rho : \mathbb{R} \to \mathbb{R}$ is as in Remark 3.16.

Example 3.24. *Let (Ω, ψ_Ω) be a weighted space and define (X, τ_X) as the dual of $\mathcal{B}_{\psi_\Omega}(\Omega)$ equipped with the weak-$*$-topology, which can be identified by the Riesz representation theorem in [DT10, Theorem 2.4] with the space of finite signed Radon measures $\mathcal{M}$ consisting of $\mu : \mathcal{F} \to \mathbb{R}$ satisfying $\int_\Omega \psi_\Omega(\omega) d\mu(\omega) < \infty$, where $\mathcal{F}$ denotes the Borel σ-algebra of Ω. For $Z = Y = \mathbb{R}$, an additive family is by Lemma 3.23 given by*

$$\mathcal{H} = \left\{ h : \mathcal{M} \to \mathbb{R};\, h(\mu) = \int_\Omega f(\omega) d\mu(\omega) + b,\, f \in \mathcal{B}_{\psi_\Omega}(\Omega),\, b \in \mathbb{R} \right\},$$

whereas the activation family $\mathcal{R}$ consists of activation functions $\rho : \mathbb{R} \to \mathbb{R}$ described in Remark 3.16. These abstract neural networks $\varphi : \mathcal{M} \to \mathbb{R}$ could be used to learn a (probability) measure over (Ω, ψ_Ω).

These results give a first overview of the broad variety of abstract neural networks and their applications. In the following, we derive sufficient conditions for universal approximation of abstract neural networks.

3.3.3 Universal Approximation of Abstract Neural Networks

Universal approximation was first proven for neural networks between Euclidean spaces in [Cyb89] and in [Hor91] using functional analytical arguments including the Hahn-Banach theorem and Fourier theory. Later, other universal approximation results were found for different activation functions, such as in [Pin99], whereas other results are about the approximation rates of neural networks, e.g. [Bar93], [Can98] and [BGKP19]. In the following, we generalize the universal approximation property of abstract neural networks by lifting the finite dimensional universal approximation results to this infinite dimensional setting.

For this purpose, we still assume that the input space (X, ψ) is given by a weighted space, the hidden layer space (Z, ψ_Z) is a complete weighted topological vector space, whose topology τ_Z is locally convex, and the output space (Y, τ_Y) is given by a complete locally convex topological vector space. In the following, we start with the main universal approximation result, which can be used to lift the universal approximation property of real-valued neural networks defined on $\mathbb{R}^n$ to higher and infinite dimensional spaces.

Definition 3.25. *An architecture $(\mathcal{H}, \mathcal{R}, \mathcal{L})$ between X and Y is* compatible *with another architecture $(\mathcal{G}, \mathcal{R}, \mathcal{K})$ between Z and Z if*

(i) for every $g \in \mathcal{G}$ and $h \in \mathcal{H}$, it follows that $g \circ h \in \mathcal{H}$,

(ii) for every $h \in \mathcal{H}$, it holds that $\sup_{x \in X} \frac{\psi_Z(h(x))}{\psi(x)} < \infty$,

(iii) the set $\bigcup_{\gamma \in \Gamma,\, z \in Z} \{\gamma(z)\}$ is dense in Y, where the set $\Gamma \subset L(Z;Y)$ is given by $\Gamma = \{\gamma \in L(Z;Y);\, \forall \alpha \in \mathcal{K} : \gamma \circ \alpha \in \mathcal{L}\}$.

Remark 3.26. *The first condition in Definition 3.25 is satisfied for every additive family $\mathcal{H}$ between X and Z if the additive family $\mathcal{G}$ between Z and Z admits the simple form $\mathcal{G} = \{Z \ni z \mapsto \lambda z + b \in Z;\, \lambda \in \mathbb{N},\, b \in Z\}$, which might not be the case for more general additive families $\mathcal{G}$. The second condition ensures that the value of a hidden layer map in $\mathcal{H}$ grows under the weight $\psi_Z : Z \to]0, \infty[$ at most like $\psi : X \to]0, \infty[$, while the third condition guarantees that the variety of the linear readout $\mathcal{K}$ is transferred to the linear readout $\mathcal{L}$.*

Theorem 3.27 (Universal Approximation on $\mathcal{B}_\psi(X;Y)$)**.** *Assume that the architecture $(\mathcal{H}, \mathcal{R}, \mathcal{L})$ between X and Y is compatible with another architecture $(\mathcal{G}, \mathcal{R}, \mathcal{K})$ between Z and Z. If $\mathcal{NN}_{Z,Z}^{\mathcal{G},\mathcal{R},\mathcal{K}}$ is $\tau^0_{\psi_Z}$-dense in $\mathcal{B}_{\psi_Z}(Z;Z)$, then $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is τ^0_ψ-dense in $\mathcal{B}_\psi(X;Y)$, i.e. for every $f \in \mathcal{B}_\psi(X;Y)$, $\varepsilon > 0$ and $q \in \mathcal{Q}$ there exists some $\varphi \in \mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ such that*

$$q(f - \varphi) = \sup_{x \in X} \frac{p(f(x) - \varphi(x))}{\psi(x)} < \varepsilon,$$

where $p \in \mathcal{P}$ denotes the seminorm correponding to $q \in \mathcal{Q}$.

Proof. Let $\mathcal{A} = \{x \mapsto \lambda_1 \cos(z_1^*(h_1(x))) + \lambda_2 \sin(z_2^*(h_2(x)));\, \lambda_i \in \mathbb{R},\, z_i^* \in Z^*,\, h_i \in \mathcal{H}\}$ be a subalgebra of $C_b^0(X) \subset \mathcal{B}_\psi(X)$ and introduce the vector subspace $\mathcal{W} = \operatorname{span}\{x \mapsto a(x)\gamma(z_0);\, a \in \mathcal{A},\, \gamma \in \Gamma,\, z_0 \in Z\}$ of $\mathcal{B}_\psi(X;Y)$, which is by definition an $\mathcal{A}$-submodule, where $\Gamma := \{\gamma \in L(Z;Y);\, \forall \alpha \in \mathcal{K} : \gamma \circ \alpha \in \mathcal{L}\}$.

Now, we show that $\mathcal{W}$ lies in the τ^0_ψ-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$, for which we assume that $\mathcal{P}_Z$ is a family of seminorms generating the topology τ_Z of Z, which exists by Theorem 1.25. Fix some $\varepsilon > 0$ and $p \in \mathcal{P}$, and let $h \in \mathcal{H}$ as well as $\gamma \in \Gamma$, for which we define $C_1 := \sup_{x \in X} \psi_Z(h(x))/\psi(x)$ and the operator seminorm $C_2 := \sup_{z \in Z,\, p_Z(z) \leqslant 1} p(\gamma(z))$, for some $p_Z \in \mathcal{P}_Z$ existing by continuity of $\gamma \in L(Z;Y)$ such that $\sup_{z \in Z} p(\gamma(z)) \leqslant C_2 \sup_{z \in Z} p_Z(z)$. Using

that $\mathcal{NN}_{Z,Z}^{\mathcal{G},\mathcal{R},\mathcal{K}}$ is $\tau_{\psi_Z}^0$-dense in $\mathcal{B}_{\psi_Z}(Z;Z)$, there exists for the function $z \mapsto w'(z) := \cos(z^*(z))\, z_0 \in \mathcal{B}_{\psi_Z}(Z;Z)$, with $z^* \in Z^*$ and $z_0 \in Z$, some $\varphi' \in \mathcal{NN}_{Z,Z}^{\mathcal{G},\mathcal{R},\mathcal{K}}$ such that

$$\sup_{z\in Z} \frac{p_Z(w'(z) - \varphi'(z))}{\psi_Z(z)} < \frac{\varepsilon}{C_1 C_2}. \tag{3.3.2}$$

For the function $x \mapsto w(x) := \cos(z^*(h(x)))\, \gamma(z_0) \in \mathcal{W}$, we define the corresponding abstract neural network $\varphi := \gamma \circ \varphi' \circ h \in \mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$, where we use that $g \circ h \in \mathcal{H}$, for all $g \in \mathcal{G}$, such that (3.3.2) implies that

$$\begin{aligned}\sup_{x\in X} \frac{p(w(x) - \varphi(x))}{\psi(x)} &\leqslant \sup_{x\in X} \frac{\psi_Z(h(x))}{\psi(x)} \sup_{z\in Z} \frac{p(\gamma w'(z) - \gamma\varphi'(z))}{\psi_Z(z)} \\ &\leqslant C_1 C_2 \sup_{z\in Z} \frac{p_Z(w'(z) - \varphi'(z))}{\psi_Z(z)} < \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ and $p \in \mathcal{P}$ were chosen arbitrarily, it follows that $x \mapsto w(x) = \cos(z^*(h(x)))\, \gamma(z_0) \in \mathcal{W}$ belongs to the τ_ψ^0-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$, which holds analogously true for $x \mapsto \sin(z^*(h(x)))\, \gamma(z_0) \in \mathcal{W}$. Hence, we conclude by the triangle inequality that linear combinations of these functions belong to the τ_ψ^0-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$, which shows that the entire $\mathcal{A}$-submodule $\mathcal{W}$ lies in the τ_ψ^0-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$.

Finally, we apply the weighted Stone-Weierstrass in Theorem 3.11 to show that $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$. For this, we observe that $\mathcal{A}$ vanishes nowhere and consists of bounded and point separating functions such that it is by Remark 3.12 locally point separating under every $\psi_{w,p} : X \to]0,\infty[$, whereas for some $x \in X$ and $y^* \in Y^*$ with $y^*(w(x)) = 0$, for all $w \in \mathcal{W}$, it follows that $y^*(\gamma(z)) = 0$, for all $z \in Z$ and $\gamma \in \Gamma$. Using that $\bigcup_{\gamma\in\Gamma,\, z\in Z}\{\gamma(z)\}$ is by assumption dense in Y, we conclude that $y^* : Y \to \mathbb{R}$ vanishes everywhere such that $\mathcal{W}(x) = \{w(x);\, w \in \mathcal{W}\}$ is by Hahn-Banach in Theorem 1.29 dense in Y, for all $x \in X$. Hence, $\mathcal{W}$ is by the weighted Stone-Weierstrass in Theorem 3.11 τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$. However, because $\mathcal{W}$ lies by the above arguments in the τ_ψ^0-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$, $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is also τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$. □

In the following, we show particular applications of Theorem 3.27 to obtain universal approximation results between different input and output spaces, where we consider the cases of Example 3.17-3.19.

Corollary 3.28 (Example 3.17(i)). *Let $X = \mathbb{R}^n$ and $Z = Y = \mathbb{R}^d$ with weight functions $\psi(x) = \exp\left(\|x\|^2\right)$ and $\psi_Z(z) = \exp\left(\|z\|\right)$, respectively. Let $\mathcal{H} = \left\{x \mapsto Ax + b;\, A \in \mathbb{R}^{d\times n},\, b \in \mathbb{R}^d\right\}$ be the additive family, assume that $\mathcal{R} = \{\rho\}$ consists of a singleton $\rho : \mathbb{R}^d \to \mathbb{R}^d$ obtained by the componentwise application of a function $\rho_0 \in C_b^0(\mathbb{R})$ satisfying one of the conditions in Remark 3.16, and suppose that the linear readout is $\mathcal{L} = \{z \mapsto diag(\beta_1, ..., \beta_d)z;\, \beta_1, ..., \beta_d \in \mathbb{R}\}$. Then, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d}$ is τ^0_ψ-dense in $\mathcal{B}_\psi(\mathbb{R}^n;\mathbb{R}^d)$.*

Proof. Since the abstract neural networks in $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d}$ are obtained by stacking real-valued neural networks on top of each other (see Example 3.17(i)), we observe that each $\varphi \in \mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d}$ can be written as

$$\varphi(x) = \sum_{k=1}^N \beta_k \rho(h_k(x)) = \begin{pmatrix} \sum_{k=1}^N \beta_k^1 \rho_0\left(\left(a_k^1\right)^\top x + b_k^1\right) \\ \vdots \\ \sum_{k=1}^N \beta_k^d \rho_0\left(\left(a_k^d\right)^\top x + b_k^d\right) \end{pmatrix}$$
$$= \sum_{k=1}^N \sum_{i=1}^d \widetilde{\beta}_k^i \rho_0\left(\left(a_k^i\right)^\top x + b_k^i\right),$$

for $\widetilde{\beta}_k^i = \beta_k^i e_i$, with e_i denoting the ith unit vector in $\mathbb{R}^d$, such that $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d} = \mathcal{NN}^{\mathcal{H}_0,\{\rho_0\},\mathcal{L}_0}_{\mathbb{R}^n,\mathbb{R}^d}$, where $\mathcal{H}_0 = \left\{\mathbb{R}^n \ni x \mapsto a^\top x + b \in \mathbb{R};\, a \in \mathbb{R}^n,\, b \in \mathbb{R}\right\}$ and $\mathcal{L}_0 = \left\{\mathbb{R} \ni z \mapsto \beta z e_i \in \mathbb{R}^d;\, \beta \in \mathbb{R},\, i = 1, ..., d\right\}$. Now, we consider the additive family $\mathcal{G} = \{\mathbb{R} \ni z \mapsto \lambda z + c \in \mathbb{R};\, \lambda \in \mathbb{N},\, c \in \mathbb{R}\}$ on $\mathbb{R}$ with weight $\psi_{\mathbb{R}}(z) = \exp\left(|z|^2\right)$ and the linear readout $\mathcal{K} = \{\mathbb{R} \ni z \mapsto wz \in \mathbb{R};\, w \in \mathbb{R}\}$, where we use that the function $\rho_0 \in C_b^0(\mathbb{R})$ satisfies one of the conditions in Remark 3.16 such that $\mathcal{NN}^{\mathcal{G},\{\rho_0\},\mathcal{K}}_{\mathbb{R},\mathbb{R}}$ is $\tau^0_{\psi_{\mathbb{R}}}$-dense in $\mathcal{B}_{\psi_{\mathbb{R}}}(\mathbb{R})$. Hence, we only need to check that $(\mathcal{H}_0, \mathcal{R}, \mathcal{L}_0)$ is compatible with $(\mathcal{G}, \mathcal{R}, \mathcal{K})$ such that $\mathcal{NN}^{\mathcal{H}_0,\{\rho_0\},\mathcal{L}_0}_{\mathbb{R}^n,\mathbb{R}^d}$ is by Theorem 3.27 τ^0_ψ-dense in $\mathcal{B}_\psi(\mathbb{R}^n;\mathbb{R}^d)$. Indeed, it follows for every $x \mapsto h(x) = a^\top x + b \in \mathcal{H}_0$ and $g(z) = \lambda z + c \in \mathcal{G}$ that $x \mapsto (g \circ h)(x) = (\lambda a)^\top x + (\lambda b + c) \in \mathcal{H}_0$. Moreover, we observe for every $x \mapsto h(x) = a^\top x + b \in \mathcal{H}_0$ that

$$\sup_{x\in X} \frac{\psi_Z(h(x))}{\psi(x)} = \sup_{x\in X} \frac{e^{\left|a^\top x + b\right|}}{e^{\|x\|^2}} \leqslant \exp\left(|b|\right) \sup_{x\in X} \frac{e^{\|a\|\|x\|}}{e^{\|x\|^2}} < \infty.$$

In addition, we observe that $\Gamma := \{\gamma \in L(Z_0; Y);\, \forall \alpha \in \mathcal{K} : \gamma \circ \alpha \in \mathcal{L}_0\}$ coincides with $\Gamma = \{z \mapsto ze_i;\, i = 1, ..., d\}$ such that $\bigcup_{\gamma\in\Gamma,\, z\in\mathbb{R}} \{\gamma(z)\} = \mathbb{R}^d$. Hence, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d} = \mathcal{NN}^{\mathcal{H}_0,\{\rho_0\},\mathcal{L}_0}_{\mathbb{R}^n,\mathbb{R}^d}$ is by Theorem 3.27 τ^0_ψ-dense in $\mathcal{B}_\psi(\mathbb{R}^n;\mathbb{R}^d)$. □

Remark 3.29. *Corollary 3.28 is a generalization of the real-valued universal approximation results in [Cyb89] and [Hor91] to weighted spaces, where we obtain universal approximation of neural networks $\varphi : \mathbb{R}^n \to \mathbb{R}$ on compacta by choosing for a compact subset $K \subset \mathbb{R}^n$ the weight function $\psi(x) = 1$.*

Corollary 3.30 (Example 3.17(ii))**.** *Let $X = \mathbb{R}^n$, $Z = \mathbb{R}^m$ and $Y = \mathbb{R}^d$ with weight functions $\psi(x) = \exp\left(\|x\|^2\right)$ and $\psi_Z(z) = \exp(\|z\|)$, respectively. Let $\mathcal{H} = \{x \mapsto Ax + b;\, A \in \mathbb{R}^{m\times n},\, b \in \mathbb{R}^m\}$ be the additive family, assume that $\mathcal{R} = \{\rho\}$ consists of a singleton $\rho : \mathbb{R}^m \to \mathbb{R}^m$ obtained by the componentwise application of a function $\rho_0 \in C_b^0(\mathbb{R})$ satisfying one of the conditions in Remark 3.16, and suppose that the linear readout is given by $\mathcal{L} = \left\{z \mapsto Wz;\, W \in \mathbb{R}^{d\times m}\right\}$. Then, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d}$ is τ^0_ψ-dense in $\mathcal{B}_\psi(\mathbb{R}^n;\mathbb{R}^d)$.*

Proof. Similarly as in the proof of Corollary 3.28, we can write an abstract neural network $\varphi \in \mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d}$ as

$$\varphi(x) = \sum_{k=1}^{N} \beta_k \rho(h_k(x)) = \sum_{k=1}^{N}\sum_{i=1}^{m} \widetilde{\beta}^i_k \rho_0\left(\left(a^i_k\right)^\top x + b^i_k\right),$$

for $\widetilde{\beta}^i_k = \beta^i_k v_i$, where $v_i \in \mathbb{R}^d$ can be any vector. In this case, we have $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d} = \mathcal{NN}^{\mathcal{H}_0,\{\rho_0\},\mathcal{L}_0}_{\mathbb{R}^n,\mathbb{R}^d}$, with $\mathcal{H}_0 = \left\{\mathbb{R}^n \ni x \mapsto a^\top x + b \in \mathbb{R};\, a \in \mathbb{R}^n,\, b \in \mathbb{R}\right\}$ and $\mathcal{L}_0 = \left\{\mathbb{R} \ni z \mapsto zv \in \mathbb{R}^d;\, v \in \mathbb{R}^d\right\}$. Then, the proof follows along the lines of the proof of Corollary 3.28, where $\Gamma = \left\{\mathbb{R} \ni z \mapsto zv \in \mathbb{R}^d;\, v \in \mathbb{R}^d\right\}$ such that $\bigcup_{\gamma\in\Gamma,\, z\in\mathbb{R}}\{\gamma(z)\} = \mathbb{R}^d$. Hence, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}^d} = \mathcal{NN}^{\mathcal{H}_0,\{\rho_0\},\mathcal{L}_0}_{\mathbb{R}^n,\mathbb{R}^d}$ is by Theorem 3.27 τ^0_ψ-dense in $\mathcal{B}_\psi(\mathbb{R}^n;\mathbb{R}^d)$. □

Corollary 3.31 (Example 3.18)**.** *Let (X,ψ) be a weighted space, let $(\mathbb{R},\psi_\mathbb{R})$ be the hidden layer space and let $Y = \mathbb{R}$ be the output space. Assume that $\mathcal{H}$ is an additive family with $\sup_{x\in X}\psi_\mathbb{R}(h(x))/\psi(x) < \infty$, for all $h \in \mathcal{H}$, let $\mathcal{R} = \{\rho\}$ consist of a singleton $\rho \in C_b^0(\mathbb{R})$ satisfying one of the conditions in Remark 3.16, and suppose that the linear readout is $\mathcal{L} = \{z \mapsto wz;\, w \in \mathbb{R}\}$. Then, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,\mathbb{R}}$ is τ^0_ψ-dense in $\mathcal{B}_\psi(X)$.*

Proof. For the additive family $\mathcal{G} = \{\mathbb{R} \ni z \mapsto \lambda z + c \in \mathbb{R};\, \lambda \in \mathbb{N},\, c \in \mathbb{R}\}$ and the linear readout $\mathcal{K} = \{\mathbb{R} \ni z \mapsto wz \in \mathbb{R};\, w \in \mathbb{R}\}$, it follows by Remark 3.16 that $\mathcal{NN}^{\mathcal{G},\mathcal{R},\mathcal{K}}_{\mathbb{R},\mathbb{R}}$ is $\tau^0_{\psi_\mathbb{R}}$-dense in $\mathcal{B}_{\psi_\mathbb{R}}(\mathbb{R})$. Then, for $z \mapsto g(z) := \lambda z + b \in \mathcal{G}$ and $h \in \mathcal{H}$,

it follows that $x \mapsto (g \circ h)(x) = \lambda h(x) + b \in \mathcal{H}$ and $\sup_{x \in X} \psi_{\mathbb{R}}(h(x))/\psi(x) < \infty$, and we have $\Gamma = \{z \mapsto wz;\, w \in \mathbb{R}\}$, which implies $\bigcup_{\gamma \in \Gamma,\, z \in \mathbb{R}} \{\gamma(z)\} = \mathbb{R}$, such that $\mathcal{NN}_{X,\mathbb{R}}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is by Theorem 3.27 τ_ψ^0-dense in $\mathcal{B}_\psi(X)$. □

Corollary 3.32 (Example 3.19)**.** *Let (X, ψ) be a weighted space, let $(\mathbb{R}^m, \psi_{\mathbb{R}^m})$ be the hidden layer space and let (Y, τ_Y) be a complete locally convex topological vector space. Let $\mathcal{H}$ be an additive family with $\sup_{x \in X} \psi_{\mathbb{R}^m}(h(x))/\psi(x) < \infty$, for all $h \in \mathcal{H}$, assume that $\mathcal{R} = \{\rho\}$ consists of a singleton $\rho : \mathbb{R}^m \to \mathbb{R}^m$ obtained by componentwise application of $\rho_0 \in C_b^0(\mathbb{R})$ satisfying one condition in Remark 3.16, and let $\mathcal{L} = \operatorname{span}\left\{\mathbb{R}^m \ni z \mapsto yw^\top z \in Y;\, w \in \mathbb{R}^m,\, y \in Y\right\}$. Then, $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X; Y)$.*

Proof. Similarly as in the previous proofs, we can write an abstract neural network $\varphi \in \mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ as

$$\varphi(x) = \sum_{k=1}^{N} \beta_k \rho(h_k(x)) = \sum_{k=1}^{N} \sum_{i=1}^{m} \widetilde{\beta}_k^i \rho_0\left(h_k^i(x)\right),$$

for $\widetilde{\beta}_k^i = \beta_k^i y_i$, where $y_i \in Y$ and h_k^i is the ith component. Then, $\mathcal{NN}_{\mathbb{R}^n,\mathbb{R}^d}^{\mathcal{H},\mathcal{R},\mathcal{L}} = \mathcal{NN}_{\mathbb{R}^n,\mathbb{R}^d}^{\mathcal{H}_0,\{\rho_0\},\mathcal{L}_0}$, where $\mathcal{H}_0 = \left\{X \ni x \mapsto h^i(x) \in \mathbb{R};\, h \in \mathcal{H},\, i = 1, ..., d\right\}$ and $\mathcal{L}_0 = \{\mathbb{R} \ni z \mapsto zy \in Y;\, y \in Y\}$. Then, the proof follows along the lines of the proof of Corollary 3.28, where we use $\Gamma = L(\mathbb{R}; Y)$ such that $\bigcup_{\gamma \in \Gamma,\, z \in \mathbb{R}} \{\gamma(z)\} = Y$. Hence, $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}} = \mathcal{NN}_{X,Y}^{\mathcal{H}_0,\{\rho_0\},\mathcal{L}_0}$ is by Theorem 3.27 τ_ψ^0-dense in $\mathcal{B}_\psi(X, Y)$. □

Corollary 3.28-3.32 are the universal approximation results of the four special learning problems presented in Example 3.17-3.19, where the hidden layer is always finite dimensional. Similarly as in Remark 3.29, the weight function $\psi : X \to]0, \infty[$ can be chosen as $\psi(x) = 1$ to obtain the universal approximation results on compacta within the space $C^0(X; Y)$ of continuous and vector-valued functions.

The arguments in Theorem 3.27 can be iterated to obtain universal approximation of deep abstract neural networks with finitely many hidden layers, which we show in the following for the particular case of an abstract neural network with two hidden layers. Let (Z_1, ψ_{Z_1}) and (Z_2, ψ_{Z_2}) be the weighted locally convex topological vector spaces representing the two hidden layers, for which we assume that there are architectures $(\mathcal{G}_1, \mathcal{R}_1, \mathcal{K}_1)$ between Z_1 and Z_1 and $(\mathcal{G}_2, \mathcal{R}_2, \mathcal{K}_2)$ between Z_2 and Z_2 such that $\mathcal{NN}_{Z_1,Z_1}^{\mathcal{G}_1,\mathcal{R}_1,\mathcal{K}_1}$ is $\tau_{\psi_{Z_1}}^0$-dense in $\mathcal{B}_{\psi_{Z_1}}(Z_1; Z_1)$ and $\mathcal{NN}_{Z_2,Z_2}^{\mathcal{G}_2,\mathcal{R}_2,\mathcal{K}_2}$ is $\tau_{\psi_{Z_2}}^0$-dense in $\mathcal{B}_{\psi_{Z_2}}(Z_2; Z_2)$.

For simplicity, we assume that the additive families $\mathcal{G}_1$ and $\mathcal{G}_2$ are given by $\mathcal{G}_{1,2} = \{Z_{1,2} \ni z_{1,2} \mapsto \lambda z_{1,2} + b \in Z_{1,2};\, \lambda \in \mathbb{N},\, b \in Z_{1,2}\}$. In addition, we suppose that $\mathcal{L}_1 \subset L(Z_1; Z_2)$ and $\mathcal{L}_2 \subset L(Z_2; Y)$ are linear readouts. Then, by showing that $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is an additive family between X and Z_2, we can apply Theorem 3.27 again to show that abstract neural networks with two hidden layers have the universal approximation property.

Theorem 3.33 (Universal Approximation of Deep ANN)**.** *Let $(\mathcal{G}_1, \mathcal{R}_1, \mathcal{K}_1)$ and $(\mathcal{G}_2, \mathcal{R}_2, \mathcal{K}_2)$ be two architectures between Z_1 and Z_1 or Z_2 and Z_2, respectively, such that $\mathcal{NN}_{Z_1,Z_1}^{\mathcal{G}_1,\mathcal{R}_1,\mathcal{K}_1}$ is $\tau^0_{\psi_{Z_1}}$-dense in $\mathcal{B}_{\psi_{Z_1}}(Z_1; Z_1)$ and $\mathcal{NN}_{Z_2,Z_2}^{\mathcal{G}_2,\mathcal{R}_2,\mathcal{K}_2}$ is $\tau^0_{\psi_{Z_2}}$-dense in $\mathcal{B}_{\psi_{Z_2}}(Z_2; Z_2)$. Assume that $\mathcal{H}$ is an additive family between X and Z_1, which satisfies $\sup_{x\in X} \psi_Z(h(x))/\psi(x) < \infty$, for all $h \in \mathcal{H}$, and suppose that the linear readouts $\mathcal{L}_1 \subset L(Z_1; Z_2)$ and $\mathcal{L}_2 \subset L(Z_2; Y)$ are such that $\bigcup_{\gamma_1\in\Gamma_1,\, z_1\in Z_1}\{\gamma_1(z_1)\}$ is dense in Z_2 and $\bigcup_{\gamma_2\in\Gamma_2,\, z_2\in Z_2}\{\gamma_2(z_2)\}$ is dense in Y, where*

$$\Gamma_1 := \{\gamma_1 \in L(Z_1; Z_2);\, \forall \alpha_1 \in \mathcal{K}_1 : \gamma_1 \circ \alpha_1 \in \mathcal{L}_1\},$$
$$\Gamma_2 := \{\gamma_2 \in L(Z_2; Y);\, \forall \alpha_2 \in \mathcal{K}_2 : \gamma_2 \circ \alpha_2 \in \mathcal{L}_2\}.$$

If $\sup_{x\in X} \psi_{Z_2}(\varphi(x))/\psi(x) < \infty$, for all $\varphi \in \mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$, the set of ANNs

$$\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1,\mathcal{R}_2,\mathcal{L}_2} := \left\{ x \mapsto \sum_{k=1}^{N_2} \beta_k \rho_k \left(\sum_{l=1}^{N_1} \beta_{kl} \rho_{kl}(h_{kl}(x)) \right);\ \begin{matrix} h_{kl} \in \mathcal{H} \\ \rho_{kl} \in \mathcal{R}_1,\, \rho_k \in \mathcal{R}_2 \\ \beta_{kl} \in \mathcal{L}_1,\, \beta_k \in \mathcal{L}_2 \end{matrix} \right\}$$

with two hidden layers is τ^0_ψ-dense in $\mathcal{B}_\psi(X; Y)$.

Proof. Since the architecture $(\mathcal{H}, \mathcal{R}_1, \mathcal{L}_1)$ between X and Z_2 is compatible with the architecture $(\mathcal{G}_1, \mathcal{R}_1, \mathcal{K}_1)$ between Z_1 and Z_1, $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is by Theorem 3.27 τ^0_ψ-dense in $\mathcal{B}_\psi(X; Z_2)$. Now, we assume by contradiction that $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is not point separating or does not contain the constant functions such that there exist distinct points $x_1, x_2 \in X$ with $\varphi(x_1) = \varphi(x_2)$, for all $\varphi \in \mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$, or there exists some proper closed subspace $E \subset Y$ such that $\operatorname{span}\{\beta\rho(b);\, \beta \in \mathcal{L}_1,\, \rho \in \mathcal{R}_1,\, b \in Z_1\} \subset E$. In both cases, it follows that the τ^0_ψ-closure of $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is a proper closed vector subspace of $\mathcal{B}_\psi(X; Z_2)$, which contradicts the fact that $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is τ^0_ψ-dense in $\mathcal{B}_\psi(X; Z)$, such that $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ must be point separating and must contain the constant functions. In addition, because $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is closed under addition, it follows that $\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}$ is an additive family between X and Z_2. Then, the architecture $\left(\mathcal{NN}_{X,Z_2}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1}, \mathcal{R}_2, \mathcal{L}_2\right)$ between X and Y is by assumption compatible

with the architecture $(\mathcal{G}_2, \mathcal{R}_2, \mathcal{K}_2)$ between Z_2 and Z_2 such that by applying Theorem 3.27, it follows that $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R}_1,\mathcal{L}_1,\mathcal{R}_2,\mathcal{L}_2}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$. □

By iterating the argument, Theorem 3.33 can be generalized to deep abstract neural networks with finitely many hidden layers. In addition, the argument can be generalized to $L^p(\Omega)$-weighted vector-valued functions, where $\mu : \mathcal{F} \to [0,\infty]$ is a possibly non-finite, positive Radon measure on X, which is finite on every compact subset $K \subset X$ and therefore in particular on every preimage $K_R = \psi^{-1}(]0,R])$, for all $R > 0$.

So far, we have lifted the universal approximation property from real-valued neural networks, where the hidden layer space (Z, ψ_Z) was always finite dimensional. If (Z, ψ_Z) becomes infinite dimensional and $\mathcal{R} \subset \mathcal{B}_{\psi_Z}(Z;Z)$ consists of several activation functions, $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ has under some additional assumptions the universal approximation property if for every hidden layer map $h \in \mathcal{H}$ there exists a direction given by $z_0 \in Z\backslash\{0\}$, with $x \mapsto j(h(x))z_0 \in \mathcal{H}$, for some $j \in Z^*$, and there exist some $\rho \in \mathcal{R}$ and $z^* \in Z^*$ such that the function $\mathbb{R} \ni t \mapsto z^*(\rho(tz_0)) \in \mathbb{R}$ satisfies one of the conditions in Remark 3.16.

3.4 Smooth Neural Networks in Infinite Dimension

In the previous section, we have considered the approximation of a given function by an abstract neural network, while for certain applications the simultaneous approximation of the function and its derivatives might be of particular interest, e.g. in Physics, where the velocity along a path should also be approximated arbitrarily well. For the finite dimensional case, [HSW90] has proved the universal approximation property including the derivatives of conventional neural networks between Euclidean spaces. For the vector-valued case in this weighted setting, we first introduce continuously differentiable functions between the input space (X, τ_X) with values in the output space (Y, τ_Y).

3.4.1 Differentiable Lipschitz Functions on Weighted Space

For the input space (X, τ_X), we introduce the notion of a manifold over convenient vector spaces. A *chart* (U, ϕ) on X is a bijection $\phi : U \to \phi(U) \subset E$ from a subset $U \subset X$ onto a c^∞-open subset $\phi(U)$ of a convenient vector space (E, τ_E). A family of charts $(U_i, \phi_i)_{i\in I}$ with $\phi_i(U_i) \subset E_i$, for convenient vector spaces $(E_i, \tau_{E_i})_{i\in I}$, is a *$C^\infty$-atlas over* $(E_i, \tau_{E_i})_{i\in I}$ if $(U_i)_{i\in I}$ forms an open cover of X and $\phi_i \circ \phi_j^{-1} : \phi_j(U_i \cap U_j) \to E_i$ is smooth, for all $i, j \in I$. If such a C^∞-atlas exists for X, then X is called a *C^∞-manifold over* $(E_i, \tau_{E_i})_{i\in I}$.

For this section, we assume that the input space (X, τ_X) is a weighted space (X, ψ), which is also a smoothly Hausdorff C^∞-manifold over convenient vector spaces $(E_i, \tau_{E_i})_{i\in I}$ with atlas $(U_i, \phi_i)_{i\in I}$, where *smoothly Hausdorff* means that $C^\infty(X)$ separates points of X such that (X, τ_X) is by [KM97, Section VI.27.4] topologically Hausdorff. Moreover, if the convenient vector spaces $(E_i, \tau_{E_i})_{i\in I}$ are finite dimensional, we recover the classical definition of a manifold over finite dimensional spaces. On the other hand, the output space (Y, τ_Y) is also a convenient vector space, with topology τ_Y generated by the family of seminorms $\mathcal{P}$.

For $k \in \mathbb{N}_{\geqslant 1}$, let $\mathcal{L}\mathrm{ip}^k(X;Y)$ be the vector space of functions $f : X \to Y$ such that $f \circ c \in \mathcal{L}\mathrm{ip}^k(\mathbb{R};Y)$, for all $c \in C^\infty(\mathbb{R};X)$, which is equivalent to $f \circ \phi_i^{-1} \in \mathcal{L}\mathrm{ip}^k(\phi_i(U_i);Y)$, for all $i \in I$, see [KM97, Lemma VI.27.17]. On $\mathcal{L}\mathrm{ip}^k(X;Y)$, we define the initial locally convex topology τ_c^k of C^k-convergence on compacts generated by

$$\left\{ f \mapsto \max_{j=0,\dots,k} \sup_{\substack{t\in\mathbb{R} \\ c(t)\in K}} p\left((f\circ c)^{(j)}(t)\right);\ \begin{array}{c} p\in\mathcal{P},\ c\in C^\infty(\mathbb{R};X) \\ K\subset X \text{ compact} \end{array} \right\},$$

which can be equivalently characterized by seminorms of the same form, but with $c \in C^\infty(\mathbb{R};U_i)$ and $i \in I$. Moreover, this topology τ_c^k is equivalent to the C^k-topology in Section 1.2.2 if X is finite dimensional (see [Bom67]). In addition, we define $\mathcal{L}\mathrm{ip}_b^k(X;Y)$ as the vector subspace of bounded $\mathcal{L}\mathrm{ip}^k$-functions in $\mathcal{L}\mathrm{ip}^k(X;Y)$ with bounded derivatives, which we equip with the initial locally convex topology τ_u^k of uniform C^k-convergence generated by $f \mapsto \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} p\left((f\circ c)^{(j)}(t)\right)$, for $p \in \mathcal{P}$ and $c \in C^\infty(\mathbb{R};X)$.

Using the admissible weight function $\psi : X \to]0,\infty[$, we define the weighted locally convex topology τ_ψ^k generated by

$$\mathcal{Q} = \left\{ f \mapsto \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{p\left((f\circ c)^{(j)}(t)\right)}{\psi(c(t))};\ p\in\mathcal{P},\ c\in C^\infty(\mathbb{R};X) \right\}, \tag{3.4.1}$$

which can be equivalently characterized by seminorms of the same form, but with $c \in C^\infty(\mathbb{R};U_i)$ and $i \in I$. Moreover, we observe that τ_ψ^k is a weaker topology on $\mathcal{L}\mathrm{ip}_b^k(X;Y)$ than τ_u^k.

Definition 3.34. *We define the* weighted space of vector-valued and $\mathcal{L}\mathrm{ip}^k$-functions $\mathcal{L}ip_\psi^k(X;Y)$ *as the c^∞-completion of $(\mathcal{L}ip_b^k(X;Y), \tau_\psi^k)$, which is by definition a convenient vector space with topology τ_ψ^k generated by the family of seminorms $\mathcal{Q}$ in* (3.4.1)*. If $Y = \mathbb{R}$, we write $\mathcal{L}ip_\psi^k(X)$ instead of $\mathcal{L}ip_\psi^k(X;\mathbb{R})$.*

Hereby, the growth of the functions in $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$ as well as the growth of their derivatives is controlled by the growth of $\psi : X \to]0,\infty[$, where the Lipschitz constant of the kth derivative is not included.

3.4.2 Weighted Nachbin Theorem with Convenient Calculus

In this section, we generalize the classical Nachbin result in Theorem 1.37 to this weighted setting, where we assume that the input space (X,ψ) is a weighted space, which is a C^∞-manifold over finite dimensional spaces $(E_i,\tau_{E_i})_{i\in I}$. First, we consider the real-valued case with $Y=\mathbb{R}$ and later extend the result to the vector-valued case with convenient vector space (Y,τ_Y) as output space, where we use polynomial algebras as in [Pro77]. Similar results for differentiable functions over an infinite dimensional Banach space have been found in [PG76] and [AP80], for the topology of compact convergence and the compact-open topology, respectively. In addition, [Nac91] showed a weighted Nachbin result for polynomials over $\mathbb{R}^d$.

Definition 3.35. *A subalgebra $\mathcal{A}\subset\mathcal{L}ip^k_\psi(X)$ is called* locally point separating of order k (under $\psi : X \to]0,\infty[$) *if there exists a point separating vector subspace $\widetilde{\mathcal{A}}\subset\mathcal{A}$ such that*

(i) for every $\widetilde{a}\in\widetilde{\mathcal{A}}$ and $c\in C^\infty(\mathbb{R};X)$, it holds that

$$\lim_{R\to\infty}\max_{j=1,\dots,k}\sup_{\substack{t\in\mathbb{R}\\ c(t)\in X\setminus K_R}}\frac{\left(1+\left|(\widetilde{a}\circ c)^{(j)}(t)\right|^k\right)\exp\left(|\widetilde{a}(c(t))|\right)}{\psi(c(t))}=0, \quad (3.4.2)$$

with the convention that the supremum over the empty set is set to be zero, where $K_R:=\psi^{-1}(]0,R])=\{x\in X;\,\psi(x)\leqslant R\}$ denotes the compact preimage,

(ii) for every $i\in I$ there exists some $m\in\mathbb{N}_{\geqslant 1}$ and $\widetilde{\eta}_1,\dots,\widetilde{\eta}_m\in\widetilde{\mathcal{A}}$ such that the function $\widetilde{\eta}:U_i\to\mathbb{R}^m$ defined by $\widetilde{\eta}(x)=(\widetilde{\eta}_1(x),\dots,\widetilde{\eta}_m(x))^\top$, for $x\in U_i$, is an embedding, i.e. continuous and injective.

Remark 3.36. *The condition (i) guarantees that the exponential series together with the derivatives (via the upper estimate of the Bell polynomials on the right hand-side of* (3.4.6)*) converges under the weight function* $\psi : X \to]0, \infty[$*, which coincides for* $k = 0$ *with the notion of locally point separating in the non-differentiable case introduced in Definition 3.5. On the other hand, the second condition in (ii) ensures that every subset of* X *can be detected through elements of the point separating vector subspace* $\widetilde{\mathcal{A}} \subset \mathcal{A}$*.*

We start with proving the weighted Nachbin theorem over an open subset $U \subset \mathbb{R}^d$ without manifold structure, where $\psi : U \to]0, \infty[$ is an admissible weight function on U. In this case, the topology τ_c^k on $\mathcal{L}\mathrm{ip}^k(U)$ coincides with the topology of uniform convergence on compact subsets of U, whereas the weighted topology τ_ψ^k on $\mathcal{L}\mathrm{ip}_\psi^k(U)$ can be equivalently generated by the weighted norm $\|\cdot\|_{\mathcal{L}\mathrm{ip}_\psi^k(U)}$ defined by

$$\|f\|_{\mathcal{L}\mathrm{ip}_\psi^k(U)} = \max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in U} \frac{|\partial^\alpha f(x)|}{\psi(x)},$$

for $f \in \mathcal{L}\mathrm{ip}_\psi^k(U)$, where $\partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ denotes the partial derivatives, and where $\alpha \in \mathbb{N}_k^d$ denotes a multi-index, with $\mathbb{N}_k^d = \{\alpha \in \mathbb{N}^d;\, |\alpha| \leqslant k\}$. Moreover, we observe that if $U \subset \mathbb{R}^d$ is bounded, τ_ψ^k is equivalent to the topology induced by $\|\cdot\|_{C^k(\overline{U})}$.

Remark 3.37. *If* $X := U \subset \mathbb{R}^d$ *is open with weight function* $\psi : U \to]0, \infty[$*, we observe by choosing curves of the form* $c(t) = xt^j$ *in* (3.4.2)*, for some* $x \in X$ *and* $j = 0, ..., k$*, that the condition* (3.4.2) *implies that*

$$\lim_{R \to \infty} \max_{1 \leqslant |\alpha| \leqslant k} \sup_{x \in U \setminus K_R} \frac{\left(1 + |\partial^\alpha \widetilde{a}(x)|^k\right) \exp\left(|\widetilde{a}(x)|\right)}{\psi(x)} = 0, \tag{3.4.3}$$

which we use as definition of locally point separating of order k *if* $X := U \subset \mathbb{R}^d$ *is finite dimensional. Hereby, the supremum in* (3.4.3) *can also be taken over* $U \setminus B_R(0)$ *instead of* $U \setminus K_R$*, where* $B_R(0) := \{x \in \mathbb{R}^d;\, \|x\| < R\}$*. In particular, the polynomials in* $\mathrm{Pol}(U)$ *are locally point separating of order* k *under the weight* $\psi(x) = \exp\left(\|x\|^2\right)$ *with point separating subspace* $\widetilde{\mathcal{A}} = \{U \ni x \mapsto wx_i \in \mathbb{R};\, w \in \mathbb{R},\, i = 1, ..., d\}$*.*

In addition, we need the following multivariate Leibniz formula and Faà di Bruno formula, which generalize their one-dimensional versions in [Apo67, Example 5.11.4] and [Faà55] as well as [Faà57], respectively.

Lemma 3.38 (Multivariate Leibniz Rule)**.** *For $U \subset \mathbb{R}^d$ open, let $f, g \in C^k(U)$. Then, we have*

$$\partial^\alpha (f \cdot g)(x) = \sum_{0 \leqslant \beta \leqslant \alpha} \binom{\alpha}{\beta} \partial^\beta f(x) \partial^{\alpha-\beta} g(x),$$

for all $x \in U$ and $\alpha \in \mathbb{N}_k^d = \left\{\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}^d;\ |\alpha| := \alpha_1 + ... + \alpha_d \leqslant k\right\}$, where $\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha-\beta)!}$ with $\alpha! = \alpha_1! \cdots \alpha_d!$, and where $\beta \leqslant \alpha$ if and only if $\beta_i \leqslant \alpha_i$, for all $i = 1, ..., d$. In addition, it holds that

$$\left|\partial^\alpha (f \cdot g)(x)\right| \leqslant 2^k \max_{0 \leqslant |\beta| \leqslant |\alpha|} \left|\partial^\beta f(x)\right| \max_{0 \leqslant |\gamma| \leqslant |\alpha|} \left|\partial^\gamma f(x)\right|, \tag{3.4.4}$$

for all $\alpha \in \mathbb{N}_k^d$ and $x \in U$.

Proof. While (3.38) can be shown by induction on $|\alpha| = 0, ..., k$ with the product rule, the inequality (3.4.4) follows by estimating (3.38) pointwise from above, where $\sum_{0 \leqslant \beta \leqslant \alpha} \binom{\alpha}{\beta} = 2^{|\alpha|} \leqslant 2^k$, for all $\alpha \in \mathbb{N}_k^d$. □

Lemma 3.39 (Multivariate Faà di Bruno Formula)**.** *For $U \subset \mathbb{R}^d$ and $V \subset \mathbb{R}^m$ open, let $f \in C^k(V)$ and $g \in C^k(U; \mathbb{R}^m)$ with $g(U) \subset V$. Then, it follows that*

$$\partial^\alpha (f \circ g)(x) = \sum_{1 \leqslant |\beta| \leqslant |\alpha|} \partial^\beta f(g(x)) \sum_{s=1}^{|\alpha|} \sum_{(k;l) \in A_s(\alpha,\beta)} \alpha! \prod_{j=1}^{s} \frac{\prod_{i=1}^m \left(\partial^{l_j} g_i(x)\right)^{k_{ji}}}{k_j! \left(l_j!\right)^{|k_j|}}, \tag{3.4.5}$$

for all $\alpha \in \mathbb{N}_k^d$ with $|\alpha| \geqslant 1$ and $x \in U$, where

$$A_s(\alpha, \beta) = \left\{(k_1, ..., k_s; l_1, ..., l_s) \in \mathbb{N}^{sm} \times \mathbb{N}^{sd};\ \begin{matrix} |k_i| > 0,\ 0 \prec l_1 \prec ... \prec l_s \\ \sum_{j=1}^s k_j = \beta,\ \sum_{j=1}^s |k_j| l_j = \alpha \end{matrix}\right\},$$

with $l \prec \widetilde{l}$ if and only if either $|l| < |\widetilde{l}|$, or $|l| = |\widetilde{l}|$ with $l_1 = \widetilde{l}_1$, ..., $l_{i-1} = \widetilde{l}_{i-1}$ and $l_i < \widetilde{l}_i$, for some $i = 1, ..., d$. In addition, it holds that

$$\left|\partial^\alpha (f \circ g)(x)\right| \leqslant C_{k,m,d} \max_{1 \leqslant |\beta| \leqslant |\alpha|} \left|\partial^\beta f(g(x))\right| \left(1 + \max_{1 \leqslant |\gamma| \leqslant |\alpha|} \max_{i=1,...,m} \left|\partial^\gamma g_i(x)\right|^k\right), \tag{3.4.6}$$

for all $x \in U$ and $1 \leqslant |\alpha| \leqslant k$, where the constant $C_{k,m,d} := 2\,(k!)^{d+m}\, k^{k(d+m)}$ does not depend on $x \in U$, $f \in C^k(V)$ and $g \in C^k(U; \mathbb{R}^m)$.

Proof. While (3.4.5) was shown in [CS96, Theorem 2.1], the inequality (3.4.6) follows by estimating (3.4.5) pointwise from above and by inserting that for every $j = 1, ..., s$, $x \in U$ and $(k; l) \in A_s(\alpha; \beta)$, it holds that

$$\prod_{i=1}^{m} \left(\partial^{l_j} g_i(x)\right)^{k_{ji}} \leqslant \max\left(1, \max_{1\leqslant|\gamma|\leqslant k} \max_{i=1,...,m} |\partial^\gamma g_i(x)|\right)^{|k_j|} \leqslant 1 + \max_{1\leqslant|\gamma|\leqslant k} \max_{i=1,...,m} |\partial^\gamma g_i(x)|^k .$$

In addition, we use that $|A_s(\alpha, \beta)| \leqslant |\alpha|^{s(d+m)}$ and $\alpha! \leqslant (k!)^d$ as well as $|\{\beta \in \mathbb{N}^m;\, 1 < |\beta| \leqslant |\alpha|\}| \leqslant (k!)^m$ such that

$$\sum_{1\leqslant|\beta|\leqslant|\alpha|} \sum_{s=1}^{|\alpha|} \sum_{(k;l)\in A_s(\alpha,\beta)} \alpha! \prod_{j=1}^{s} \frac{1}{k_j!\,(l_j!)^{|k_j|}} \leqslant (k!)^{d+m} \sum_{s=1}^{k} k^{s(d+m)} \leqslant 2\,(k!)^{d+m}\, k^{k(d+m)} = C_{k,m,d},$$

where we have inserted that $\sum_{s=1}^{k} k^{s(d+m)} = 1 \leqslant 2k^{k(d+m)}$, for $k = 1$, and $\sum_{s=1}^{k} k^{s(d+m)} = k^{d+m}\frac{k^{k(d+m)}-1}{k^{d+m}-1} \leqslant 2k^{k(d+m)}$, for all $k \geqslant 2$. □

Theorem 3.40 (Nachbin on $(\mathcal{L}\mathrm{ip}^k_\psi(U), \tau^k_\psi)$)**.** *For $U \subset \mathbb{R}^d$ open, let (U, ψ) be a weighted space. For a subalgebra $\mathcal{A} \subset \mathcal{L}ip^k_\psi(U)$, the following are equivalent.*

(i) $\mathcal{A}$ is locally point separating of order k, vanishes nowhere, and for every $x \in U$ and $v \in \mathbb{R}^d\backslash\{0\}$ there exists some $a \in \mathcal{A}$ such that $v^\top \nabla a(x) \neq 0$.

(ii) $\mathcal{A}$ is τ^k_ψ-dense in $\mathcal{L}ip^k_\psi(U)$.

Proof. In order to show that (ii) implies (i), we assume by contradiction that $\mathcal{A}$ is not locally point separating of order k, or does not vanish nowhere, or there exists some $x_3 \in \mathbb{R}^d$ and $v \in \mathbb{R}^d\backslash\{0\}$ such that $v^\top \nabla a(x_3) = 0$, for all $a \in \mathcal{A}$. If $\mathcal{A}$ is not locally point separating of order k, then $\mathcal{A}$ is either not point separating at all, or there exists a point separating vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ such that either (3.4.3) is violated, for some $\widetilde{a} \in \widetilde{\mathcal{A}}$, or for every $m \in \mathbb{N}_{\geqslant 1}$ and $\widetilde{\eta}_1, ..., \widetilde{\eta}_m \in \widetilde{\mathcal{A}}$, the function $U \ni x \mapsto \widetilde{\eta}(x) := (\widetilde{\eta}_1(x), ..., \widetilde{\eta}_m(x))^\top \in \mathbb{R}^m$ is not an embedding. In the first case, we define the proper closed subalgebra $\mathcal{A}_1 = \{f \in \mathcal{L}\mathrm{ip}^k_\psi(U);\, f(x_1) = f(x_2)\}$ of $\mathcal{L}\mathrm{ip}^k_\psi(U)$, whereas in the second case, we obtain a contradiction because the vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ contains the zero function satisfying (3.4.3). In the last case, every function $U \ni x \mapsto \widetilde{\eta}(x) := (\widetilde{\eta}_1(x), ..., \widetilde{\eta}_m(x))^\top \in \mathbb{R}^m$ can only violate the injectivity such that there exist

distinct $x_{\widetilde{\eta},1} \neq x_{\widetilde{\eta},2}$ depending on $\widetilde{\eta} : U \to \mathbb{R}^m$ with $\widetilde{\eta}(x_{\widetilde{\eta},1}) = \widetilde{\eta}(x_{\widetilde{\eta},2})$. From this, we define for every function $U \ni x \mapsto \widetilde{\eta}(x) := (\widetilde{\eta}_1(x), ..., \widetilde{\eta}_m(x))^\top \in \mathbb{R}^m$ the proper closed subalgebras $\mathcal{A}_{\widetilde{\eta}} = \big\{f \in \mathcal{L}\mathrm{ip}^k_\psi(U);\, f(x_{\widetilde{\eta},1}) = f(x_{\widetilde{\eta},2})\big\}$ of $\mathcal{L}\mathrm{ip}^k_\psi(U)$. On the other hand, if $\mathcal{A}$ does not vanish nowhere, there exists some $x_0 \in U$ with $a(x_0) = 0$, for all $a \in \mathcal{A}$, from which we define the proper closed subalgebra $\mathcal{A}_2 = \big\{f \in \mathcal{L}\mathrm{ip}^k_\psi(U);\, f(x_0) = 0\big\}$. In the third case, we define the proper closed subalgebra $\mathcal{A}_3 = \big\{f \in \mathcal{L}\mathrm{ip}^k_\psi(U);\, v^\top \nabla f(x_3) = 0\big\}$. Then, the τ^k_ψ-closure of $\mathcal{A}$ lies in one of the proper closed subalgebras $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ or $\mathcal{A}_{\widetilde{\eta}}$, for $\widetilde{\eta}_1, ..., \widetilde{\eta}_m \in \mathcal{A}$, such that the τ^k_ψ-closure of $\mathcal{A}$ is also a proper closed subalgebra of $\mathcal{L}\mathrm{ip}^k_\psi(U)$, which contradicts the assumption in (ii) such that the assertions in (i) must be satisfied.

For the other direction, it suffices to show the approximation of some $f \in \mathcal{L}\mathrm{ip}^k_b(U)$ by an element of $\mathcal{A}$ since $\mathcal{L}\mathrm{ip}^k_b(U)$ is by definition $\|\cdot\|_{\mathcal{L}\mathrm{ip}^k_\psi(U)}$-dense in $\mathcal{L}\mathrm{ip}^k_\psi(U)$. We first assume that $\mathcal{A}$ consists only of bounded functions, i.e. $\sup_{x\in U} |a(x)| < \infty$, for all $a \in \mathcal{A}$, such that by looking at (3.4.3) the assumption of $\mathcal{A}$ being locally point separating of order k reduces to $\mathcal{A}$ being point separating. Let $f \in \mathcal{L}\mathrm{ip}^k_b(U)$ be a function, fix some $\varepsilon > 0$, and define $b = \max_{0\leqslant|\alpha|\leqslant k} \sup_{x\in U} |\partial^\alpha f(x)| + \varepsilon/(6M)$, where $M := \max\Big(1, \big(\inf_{x\in U} \psi(x)\big)^{-1}\Big) > 0$. Choosing $R \geqslant 6b/\varepsilon$, we consider the compact preimage $K_R = \psi^{-1}(]0, R])$ and let $U \ni x \mapsto \eta(x) := (\eta_1(x), ..., \eta_m(x))^\top \in \mathbb{R}^m$ be an embedding with $\eta_1, ..., \eta_m \in \mathcal{A}$, where $C_0 := \max_{0\leqslant|\alpha|\leqslant k} \sup_{x\in U} \max_{i=1,...,m} |\partial^\alpha \eta_i(x)|^k / \psi(x)$. Then, there exists some $R_2 > R$ and $\theta \in C^\infty(\mathbb{R}^m)$ such that $0 \leqslant \theta(y) \leqslant 1$, for all $y \in \mathbb{R}^m$, $\theta(y) = 1$, for all $y \in \eta(K_R)$, $\theta(y) = 0$, for all $y \in \mathbb{R}^m \backslash \eta(K_{R_2})$, and

$$\max_{1\leqslant|\beta|\leqslant k} \sup_{y\in\mathbb{R}^m} \big|\partial^\beta \theta(y)\big| \leqslant \frac{\varepsilon}{2^k 6 C_{k,m,d} b (M + C_0)},$$

where $C_{k,m,d} > 0$ is the constant introduced in Lemma 3.39. Then, it follows by Lemma 3.39 that

$$\begin{aligned}\max_{1\leqslant|\alpha|\leqslant k} \sup_{x\in U} \frac{|\partial^\alpha(\theta\circ\eta)(x)|}{\psi(x)} &\leqslant C_{k,m,d} \max_{1\leqslant|\beta|\leqslant k} \sup_{y\in\mathbb{R}^m} \big|\partial^\beta\theta(y)\big| \cdots \\ &\quad \cdots \sup_{x\in U} \frac{1 + \max_{1\leqslant|\gamma|\leqslant k} \max_{i=1,...,m} |\partial^\gamma \eta_i(x)|^k}{\psi(x)} \\ &\leqslant C_{k,m,d} \frac{\varepsilon}{2^k 6 C_{k,m,d} b (M + C_0)} (M + C_0) \leqslant \frac{\varepsilon}{2^k 6 b}.\end{aligned} \tag{3.4.7}$$

Moreover, we observe that $\mathcal{A}$ is also a subalgebra of $C^k(U)$, which satisfies the assumptions of the classical Nachbin result in Theorem 1.37 such that there

exists for the compact set $K_{R_2} \subset U$ some $a \in \mathcal{A}$ with

$$\|f - a\|_{C^k(K_{R_2})} = \max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in K_{R_2}} |\partial^\alpha f(x) - \partial^\alpha a(x)| < \frac{\varepsilon}{6M}, \tag{3.4.8}$$

which implies $|\partial^\alpha a(x)| \leqslant \varepsilon/(6M) + |\partial^\alpha f(x)| \leqslant b$, for all $0 \leqslant |\alpha| \leqslant k$ and $x \in K_{R_2}$. Now, we define the function $g : \mathbb{R}^{m+1} \to \mathbb{R}$ by $g(y,z) := \theta(y)z$, for $(y,z) \in \mathbb{R}^{m+1}$, such that $g \in C^\infty(\mathbb{R}^{m+1})$ and $g(\eta(x), a(x)) = a(x)$, for all $x \in K_R$. Then, by using Lemma 3.38 and $\sup_{y \in \mathbb{R}^m} |\theta(y)| \leqslant 1$ as well as (3.4.7), it follows that

$$\begin{aligned}
&\max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in U \setminus K_R} \frac{|\partial^\alpha (g \circ (\eta \oplus a))(x)|}{\psi(x)} \\
&\quad \leqslant \max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in K_{R_2} \setminus K_R} \frac{\sum_{0 \leqslant \beta \leqslant \alpha} \binom{\alpha}{\beta} \left|\partial^\beta (\theta \circ \eta)(x)\right| \left|\partial^{\alpha-\beta} a(x)\right|}{\psi(x)} \\
&\quad \leqslant b \max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in U \setminus K_R} \frac{\sum_{0 \leqslant \beta \leqslant \alpha} \binom{\alpha}{\beta} \left|\partial^\beta (\theta \circ \eta)(x)\right|}{\psi(x)} \\
&\quad \leqslant b \max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in U \setminus K_R} \frac{|\theta(\eta(x))|}{\psi(x)} + b \max_{0 \leqslant |\alpha| \leqslant k} \sum_{\substack{0 \leqslant \beta \leqslant \alpha \\ |\beta| \geqslant 1}} \binom{\alpha}{\beta} \frac{\left|\partial^\beta (\theta \circ \eta)(x)\right|}{\psi(x)} \\
&\quad \leqslant \frac{b}{R} \underbrace{\sup_{x \in U} |\theta(\eta(x))|}_{\leqslant 1} + 2^k b \underbrace{\max_{1 \leqslant |\beta| \leqslant k} \sup_{x \in U} \frac{\left|\partial^\beta (\theta \circ \eta)(x)\right|}{\psi(x)}}_{\leqslant \varepsilon/(2^k 6b)} \leqslant \frac{\varepsilon}{3}.
\end{aligned}$$

Using this and (3.4.8), we conclude that

$$\begin{aligned}
\|f - g \circ (\eta \oplus a)\|_{\mathcal{L}\mathrm{ip}^k_\psi(U)} &= \max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in U} \frac{|\partial^\alpha f(x) - \partial^\alpha (g \circ (\eta \oplus a))(x)|}{\psi(x)} \\
&\leqslant M \sup_{x \in K_R} |\partial^\alpha f(x) - \partial^\alpha a(x)| + \max_{0 \leqslant |\alpha| \leqslant k} \sup_{x \in U \setminus K_R} \frac{|\partial^\alpha f(x)|}{\psi(x)} \\
&\quad + \max_{0 \leqslant |\beta| \leqslant k} \sup_{x \in U \setminus K_R} \frac{\left|\partial^\beta (g \circ (\eta \oplus a))(x)\right|}{\psi(x)} < M \frac{\varepsilon}{6M} + \frac{b}{R} + \frac{\varepsilon}{3} \leqslant \frac{2\varepsilon}{3}.
\end{aligned} \tag{3.4.9}$$

For $C_1 = \max_{0 \leqslant |\gamma| \leqslant k} \sup_{x \in U} |\partial^\gamma a(x)|^k / \psi(x)$, we define the constant $C_2 = 1 + C_0 + C_1$. Then, there exists by the Weierstrass approximation result with derivatives in Theorem 1.36 applied on the compact set $\widetilde{K} := \overline{\eta(U)} \times a(U) \subset$

$\mathbb{R}^{m+1}$ some polynomial $p_n \in \mathrm{Pol}(\mathbb{R}^{m+1})$ such that

$$\begin{aligned}\|g-p_n\|_{C^k(\widetilde{K})} &= \max_{0\leqslant|(\beta,j)|\leqslant k}\sup_{(y,z)\in\widetilde{K}}\left|\partial^{(\beta,j)}g(y,z)-\partial^{(\beta,j)}p_n(y,z)\right| \\ &< \frac{\varepsilon}{3MC_{k,m+1,d}C_2}.\end{aligned}$$

Using this, it follows together with $C_{k,m+1,d}\geqslant 1$ and $C_2\geqslant 1$ that

$$\begin{aligned}\sup_{x\in U}\frac{|g(\eta(x),a(x))-p_n(\eta(x),a(x))|}{\psi(x)} &\leqslant M\sup_{(y,z)\in\widetilde{K}}|g(y,z)-p_n(y,z)| \\ &< M\frac{\varepsilon}{3MC_{k,m+1,d}C_2}\leqslant\frac{\varepsilon}{3}.\end{aligned} \tag{3.4.10}$$

On the other hand, we conclude with the help of Lemma 3.39 and using $M\geqslant 1$ that

$$\begin{aligned}&\max_{1\leqslant|\alpha|\leqslant k}\sup_{x\in U}\frac{|\partial^\alpha((g-p_n)\circ(\eta\oplus a)(x))|}{\psi(x)} \\ &\quad\leqslant C_{k,m+1,d}\max_{1\leqslant|(\beta,j)|\leqslant k}\sup_{(y,z)\in\widetilde{K}}\left|\partial^{(\beta,j)}g(y,z)-\partial^{(\beta,j)}p_n(y,z)\right|\cdots \\ &\qquad\cdots\sup_{x\in U}\frac{1+\max_{1\leqslant|\gamma|\leqslant k}\max\left(\max_{i=1,\ldots,d}|\partial^\gamma\eta_i(x)|^k,|\partial^\gamma a(x)|^k\right)}{\psi(x)} \\ &\quad\leqslant C_{k,m+1,d}\frac{\varepsilon}{3MC_{k,m+1,d}C_2}C_2\leqslant\frac{\varepsilon}{3}.\end{aligned} \tag{3.4.11}$$

Combining (3.4.9) with (3.4.10) and (3.4.11), and using that $p_n\circ(\eta\oplus a)\in\mathcal{A}$ as $\mathcal{A}$ is a subalgebra, we have

$$\begin{aligned}\|f-p_n\circ(\eta\oplus a)\|_{\mathcal{L}\mathrm{ip}^k_\psi(U)} &\leqslant \|f-g\circ(\eta\oplus a)\|_{\mathcal{L}\mathrm{ip}^k_\psi(U)} \\ &\quad+\|(g-p_n)\circ(\eta\oplus a)\|_{\mathcal{L}\mathrm{ip}^k_\psi(U)}<\frac{2\varepsilon}{3}+\frac{\varepsilon}{3}=\varepsilon.\end{aligned}$$

Since $\varepsilon>0$ and $f\in\mathcal{L}\mathrm{ip}^k_b(U)$ were chosen arbitrarily, it follows that $\mathcal{A}$ is τ^k_ψ-dense in $\mathcal{L}\mathrm{ip}^k_\psi(U)$.

For the general case of a locally point separating subalgebra $\mathcal{A}\subset\mathcal{L}\mathrm{ip}^k_\psi(U)$ of order k with point separating subspace $\widetilde{\mathcal{A}}\subset\mathcal{A}$ such that (3.4.3) holds true, for all $\widetilde{a}\in\widetilde{\mathcal{A}}$, let $U\ni x\mapsto\widetilde{\eta}(x):=(\widetilde{\eta}_1(x),\ldots,\widetilde{\eta}_m(x))^\top\in\mathbb{R}^m$ with $\widetilde{\eta}_1,\ldots,\widetilde{\eta}_m\in\widetilde{\mathcal{A}}$ be an embedding. Now, we show that $x\mapsto\cos(\widetilde{a}(x))$ and $x\mapsto\sin(\widetilde{a}(x))$, with $\widetilde{a}\in\widetilde{\mathcal{A}}$, belong to the τ^k_ψ-closure of $\mathcal{A}$. Indeed, for $\widetilde{a}\in\widetilde{\mathcal{A}}$ and fixed $\varepsilon>0$, we

find by (3.4.3) some $R > 4/\varepsilon$ such that

$$\max_{1\leqslant|\alpha|\leqslant k}\sup_{x\in U\setminus K_R}\frac{\left(1+|\partial^\alpha\widetilde{a}(x)|^k\right)\exp\left(|\widetilde{a}(x)|\right)}{\psi(x)}<\frac{\varepsilon}{4C_{k,1,d}}.$$

For $C_3=\max_{0\leqslant|\alpha|\leqslant k}\sup_{x\in K_R}|\partial^\alpha\widetilde{a}(x)|$, the cosine Taylor polynomial $p_n(s)=\sum_{i=0}^n\frac{(-1)^i}{(2i)!}s^{2i}$ satisfies

$$\|\cos-p_n\|_{C^k([-C_3,C_3])}=\max_{j=0,\dots,k}\sup_{|s|\leqslant C_3}\left|\cos^{(j)}(s)-p_n^{(j)}(s)\right|\leqslant\frac{C_3^{2n-k+1}}{(2n-k+1)!},$$

for all $n\in\mathbb{N}$. Hence, by choosing $n\in\mathbb{N}$ large enough such that $\frac{C_3^{2n-k+1}}{(2n-k+1)!}\leqslant\frac{\varepsilon}{2MC_{k,1,d}\left(1+C_3^k\right)}$, it follows by using $C_{k,1,d}\geqslant 1$, $C_3\geqslant 0$ and $R\geqslant\varepsilon/4$ that

$$\begin{aligned}\sup_{x\in U}\frac{|\cos\left(\widetilde{a}(x)\right)-p_n\left(\widetilde{a}(x)\right)|}{\psi(x)}&\leqslant M\sup_{x\in K_R}\left|\cos\left(\widetilde{a}(x)\right)-p_n\left(\widetilde{a}(x)\right)\right|\\&+\sup_{x\in U\setminus K_R}\frac{|\cos\left(\widetilde{a}(x)\right)|}{\psi(x)}+\sup_{x\in U\setminus K_R}\frac{|p_n\left(\widetilde{a}(x)\right)|}{\psi(x)}<M\frac{\varepsilon}{2MC_{k,1,d}\left(1+C_3^k\right)}\\&+\frac{\sup_{x\in U}|\cos\left(\widetilde{a}(x)\right)|}{R}+\sup_{x\in U\setminus K_R}\frac{\exp\left(|\widetilde{a}(x)|\right)}{\psi(x)}\leqslant\frac{\varepsilon}{2}+\frac{1}{R}+\frac{\varepsilon}{4C_{k,1,d}}\leqslant\varepsilon.\end{aligned}$$

On the other hand, Lemma 3.39 and $\left|\cos^{(j)}(s)\right|\leqslant 1$, $\exp(|s|)\geqslant 1$ as well as $\left|p_n^{(j)}(s)\right|\leqslant\exp(|s|)$ show that

$$\begin{aligned}\max_{1\leqslant|\alpha|\leqslant k}\sup_{x\in U}\frac{|\partial^\alpha\left((\cos-p_n)\circ\widetilde{a}\right)(x)|}{\psi(x)}&\leqslant MC_{k,1,d}\|\cos-p_n\|_{C^k([-C_3,C_3])}\left(1+C_3^k\right)\\&+C_{k,1,d}\sup_{x\in U\setminus K_R}\frac{\left(1+\left|p_n^{(j)}\left(\widetilde{a}(x)\right)\right|\right)\left(1+\max_{1\leqslant|\gamma|\leqslant k}|\partial^\gamma\widetilde{a}(x)|^k\right)}{\psi(x)}\\&<MC_{k,1,d}\frac{\varepsilon}{2MC_{k,1,d}\left(1+C_3^k\right)}\left(1+C_3^k\right)\\&+2C_{k,1,d}\max_{1\leqslant|\gamma|\leqslant k}\sup_{x\in U\setminus K_R}\frac{\left(1+|\partial^\gamma\widetilde{a}(x)|^k\right)\exp\left(|\widetilde{a}(x)|\right)}{\psi(x)}\leqslant\varepsilon.\end{aligned}$$

Since $\varepsilon>0$ was chosen arbitrarily, it follows that $x\mapsto\cos\left(\widetilde{a}(x)\right)$ belongs to the τ_ψ^k-closure of $\mathcal{A}$, which holds analogously true for the function $x\mapsto\sin\left(\widetilde{a}(x)\right)$,

for all $\widetilde{a} \in \widetilde{\mathcal{A}}$. Hence, we conclude that the subalgebra

$$\mathcal{A}_{\mathrm{trig}} = \left\{ x \mapsto \lambda_1 \cos\left(\widetilde{a}_1(x)\right) + \lambda_2 \sin\left(\widetilde{a}_2(x)\right);\ \lambda_1, \lambda_2 \in \mathbb{R},\ \widetilde{a}_1, \widetilde{a}_2 \in \widetilde{\mathcal{A}} \right\}$$

of $\mathcal{L}\mathrm{ip}_\psi^k(U)$ is point separating, vanishes nowhere, and for every $x \in U$ and $v \in \mathbb{R}^d \backslash \{0\}$ there exists some $a \in \mathcal{A}_{\mathrm{trig}}$ with $v^\top \nabla a(x) \neq 0$. By using the continuous and injective function $\mathbb{R} \ni s \mapsto g(s) := (\cos(s), \sin(s), \cos(\pi s), \sin(\pi s)) \in \mathbb{R}^4$, we observe that the composition

$$x \mapsto \widetilde{\eta}_{\mathrm{trig}}(x) := \left(\cos\left(\widetilde{\eta}_j(x)\right), \sin\left(\widetilde{\eta}_j(x)\right), \cos\left(\pi\widetilde{\eta}_j(x)\right), \sin\left(\pi\widetilde{\eta}_j(x)\right)\right)_{j=1,\dots,m}$$

is also an embedding with components of $\mathcal{A}_{\mathrm{trig}}$. Hence, by applying the above arguments to $\mathcal{A}_{\mathrm{trig}}$ consisting of bounded functions with embedding $\widetilde{\eta}_{\mathrm{trig}} : X \to \mathbb{R}^{4m}$, it follows that $\mathcal{A}_{\mathrm{trig}}$ is τ_ψ^k-dense in $\mathcal{L}\mathrm{ip}_\psi^k(U)$. However, as $\mathcal{A}_{\mathrm{trig}}$ lies in the τ_ψ^k-closure of $\mathcal{A}$, $\mathcal{A}$ is also τ_ψ^k-dense in $\mathcal{L}\mathrm{ip}_\psi^k(U)$. □

Remark 3.41. *Theorem 3.40 is an extension of [Nac91, Proposition 4] to the approximation of differentiable functions defined on an open subset of* $\mathbb{R}^d$ *with subalgebra* $\mathcal{A} \subset \mathcal{L}ip_\psi^k(\mathbb{R}^d)$ *instead of* $\mathrm{Pol}(\mathbb{R}^d)$*, whereas [Nac91] allows for multiple upper semicontinuous weights* $v : X \to [0, \infty[$*, e.g.* $v(x) = 1/\psi(x)$*.*

Lemma 3.42. *Let* (X, ψ) *be a weighted space, which is a* C^∞*-manifold over finite dimensional spaces* $(E_i, \tau_{E_i})_{i \in I}$*, and define the admissible weight functions* $\psi_i : \phi_i(U_i) \to]0, \infty[$ *by* $\psi_i(x) = \psi\left(\phi_i^{-1}(x)\right)$*, for* $x \in \phi_i(U_i)$*. Then, it follows for every* $f \in \mathcal{L}ip_\psi^k(X)$ *and* $i \in I$ *that* $f \circ \phi_i^{-1} \in \mathcal{L}ip_{\psi_i}^k(\phi_i(U_i))$*.*

Proof. First, we observe that $\psi_i := \psi \circ \phi_i^{-1} : \phi_i(U_i) \to]0, \infty[$ are admissible, because every $\phi_i : U_i \to \phi_i(U_i)$ is a diffeomorphism. Now, let $f \in \mathcal{L}\mathrm{ip}_\psi^k(X)$, fix $\varepsilon > 0$, $i \in I$ and $c_i \in C^\infty(\mathbb{R}; \phi_i(U_i))$, from which we define $c := \phi_i^{-1} \circ c_i \in C^\infty(\mathbb{R}; U_i)$. Since $\mathcal{L}\mathrm{ip}_\psi^k(X)$ is defined as the c^∞-completion of $(\mathcal{L}\mathrm{ip}_b(X), \tau_\psi^k)$, there exists $g \in \mathcal{L}\mathrm{ip}_b^k(X)$ with $\max_{j=0,\dots,k} \sup_{t \in \mathbb{R}} \left|((f-g) \circ c)^{(j)}(t)\right| / \psi(c(t)) < \varepsilon$. From this, we define $g_i := g \circ \phi_i^{-1} \in \mathcal{L}\mathrm{ip}_b^k(\phi_i(U_i))$ such that

$$\max_{j=0,\dots k} \sup_{t \in \mathbb{R}} \frac{\left|\left(\left(f \circ \phi_i^{-1} - g_i\right) \circ c_i\right)^{(j)}(t)\right|}{\psi_i(c_i(t))} = \max_{j=0,\dots k} \sup_{t \in \mathbb{R}} \frac{\left|((f-g) \circ c)^{(j)}(t)\right|}{\psi(c(t))} < \varepsilon,$$

showing that $f \circ \phi_i^{-1} \in \mathcal{L}\mathrm{ip}_{\psi_i}^k(\phi_i(U_i))$, since $\mathcal{L}\mathrm{ip}_{\psi_i}^k(\phi_i(U_i))$ is defined as the c^∞-completion of $\mathcal{L}\mathrm{ip}_b^k(\phi_i(U_i))$. □

Theorem 3.43 (Nachbin on $(\mathcal{L}\mathrm{ip}^k_\psi(X), \tau^k_\psi)$)**.** *Let (X, ψ) be a weighted space, which is a C^∞-manifold over finite dimensional spaces $(E_i, \tau_{E_i})_{i\in I}$. For a subalgebra $\mathcal{A} \subset \mathcal{L}ip^k_\psi(X)$, the following are equivalent.*

(i) $\mathcal{A}$ is locally point separating of order k, vanishes nowhere, and for every $c \in C^\infty(\mathbb{R}; X)$ with $c'(0) \neq 0$ there exists some $a \in \mathcal{A}$ such that $(a \circ c)'(0) \neq 0$.

(ii) $\mathcal{A}$ is τ^k_ψ-dense in $\mathcal{L}ip^k_\psi(X)$.

Proof. In order to show that (ii) implies (i), we assume by contradiction that $\mathcal{A}$ is not locally point separating of order k, or does not vanish nowhere, or there exists some $c \in C^\infty(\mathbb{R}; X)$ with $c'(0) \neq 0$ such that $(a \circ c)'(0) = 0$, for all $a \in \mathcal{A}$. If $\mathcal{A}$ is not locally point separating of order k, then $\mathcal{A}$ is either not point separating at all, or there exists a point separating vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ such that either (3.4.3) is violated, for some $\widetilde{a} \in \widetilde{\mathcal{A}}$, or for some $i \in I$, the function $U_i \ni x \mapsto \widetilde{\eta}(x) := (\widetilde{\eta}_1(x), ..., \widetilde{\eta}_m(x))^\top \in \mathbb{R}^m$ is not an embedding, for all $m \in \mathbb{N}_{\geqslant 1}$ and $\widetilde{\eta}_1, ..., \widetilde{\eta}_m \in \widetilde{\mathcal{A}}$. In the first case, we define the proper closed subalgebra $\mathcal{A}_1 = \big\{f \in \mathcal{L}\mathrm{ip}^k_\psi(X);\, f(x_1) = f(x_2)\big\}$ of $\mathcal{L}\mathrm{ip}^k_\psi(X)$, whereas in the second case, we obtain a contradiction because the vector subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ contains the zero function satisfying (3.4.3). In the last case, every function $U_i \ni x \mapsto \widetilde{\eta}(x) := (\widetilde{\eta}_1(x), ..., \widetilde{\eta}_m(x))^\top \in \mathbb{R}^m$ can only violate the injectivity such that there exist distinct points $x_{\widetilde{\eta},1} \neq x_{\widetilde{\eta},2}$ depending on $\widetilde{\eta} : U_i \to \mathbb{R}^m$ with $\widetilde{\eta}(x_{\widetilde{\eta},1}) = \widetilde{\eta}(x_{\widetilde{\eta},2})$. From this, we define for every function $U_i \ni x \mapsto \widetilde{\eta}(x) := (\widetilde{\eta}_1(x), ..., \widetilde{\eta}_m(x))^\top \in \mathbb{R}^m$ the proper closed subalgebras $\mathcal{A}_{\widetilde{\eta}} = \big\{f \in \mathcal{L}\mathrm{ip}^k_\psi(X);\, f(x_{\widetilde{\eta},1}) = f(x_{\widetilde{\eta},2})\big\}$ of $\mathcal{L}\mathrm{ip}^k_\psi(X)$. On the other hand, if $\mathcal{A}$ does not vanish nowhere, there exists some $x_0 \in X$ with $a(x_0) = 0$, for all $a \in \mathcal{A}$, from which we define the proper closed subalgebra $\mathcal{A}_2 = \big\{f \in \mathcal{L}\mathrm{ip}^k_\psi(X);\, f(x_0) = 0\big\}$. In the third case, we define the proper closed subalgebra $\mathcal{A}_3 = \big\{f \in \mathcal{L}\mathrm{ip}^k_\psi(X);\, (f \circ c)'(0) = 0\big\}$. Then, the τ^k_ψ-closure of $\mathcal{A}$ is contained in one of the proper closed subalgebras $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ or $\mathcal{A}_{\widetilde{\eta}}$, for $\widetilde{\eta}_1, ..., \widetilde{\eta}_m \in \mathcal{A}$, such that the τ^k_ψ-closure of $\mathcal{A}$ is also a proper closed subalgebra of $\mathcal{L}\mathrm{ip}^k_\psi(X)$, which contradicts the assumption in (ii) such that the assertions in (i) must be satisfied.

For the other direction, let $f \in \mathcal{L}\mathrm{ip}^k_\psi(X)$, fix some $\varepsilon > 0$ and assume without loss of generality that $c \in C^\infty(\mathbb{R}; U_i)$, for some $i \in I$. Using Lemma 3.42, we consider the function $f \circ \phi_i^{-1} \in \mathcal{L}\mathrm{ip}^k_{\psi\circ\phi_i^{-1}}(\phi_i(U_i))$ with open subset $\phi_i(U_i) \subset E_i$ and admissible weight function $\psi_i := \psi \circ \phi_i^{-1} : \phi_i(U_i) \to]0, \infty[$, from which we define $d := \dim(\phi_i(U_i))$ and $c_i := \phi_i \circ c \in C^\infty(\mathbb{R}; \phi_i(U_i))$. Then, we observe that the subalgebra $\mathcal{A}_i := \big\{a_i := a \circ \phi_i^{-1};\, a \in \mathcal{A}\big\}$ of $\mathcal{L}\mathrm{ip}^k_{\psi_i}(\phi_i(U_i))$ is locally point separating of order k, because $\mathcal{A}$ is locally point separating of order

k with point separating subspace $\widetilde{\mathcal{A}} \subset \mathcal{A}$ such that (3.4.3) holds true under $\psi : X \to]0,\infty[$, for all $\widetilde{a} \in \widetilde{\mathcal{A}}$, from which we define the point separating vector subspace $\widetilde{\mathcal{A}}_i := \left\{\widetilde{a} \circ \phi_i^{-1};\, \widetilde{a} \in \widetilde{\mathcal{A}}\right\} \subset \mathcal{A}_i$ such that (3.4.3) holds true under $\psi_i : \phi_i(U_i) \to]0,\infty[$, for all $\widetilde{a}_i \in \widetilde{\mathcal{A}}$. Similarly, $\mathcal{A}_i$ vanishes nowhere, because for every $u_0 \in \phi_i(U_i)$ we define $x_0 = \phi_i^{-1}(u_0)$ such that there exists some $a \in \mathcal{A}$ with $a_i(u_0) = a(x_0) \neq 0$, where $a_i := a \circ \phi_i^{-1} \in \mathcal{A}$. In addition, for $u_3 \in \phi_i(U_i)$ and $v \in \mathbb{R}^d \backslash \{0\}$, let $c_{u_3,v} \in C^\infty(\mathbb{R}; \phi_i(U_i))$ be a curve with $c_{u_3,v}(0) = u_3$ and $c'_{u_3,v}(0) = v$ such that $c := \phi_i^{-1} \circ c_{u_3,v} \in C^\infty(\mathbb{R}; X)$ satisfies $c'(0) = v^\top \nabla \phi_i^{-1}(u_3) \neq 0$. Then, there exists some $a \in \mathcal{A}$ such that $v^\top \nabla a_i(u_3) = (a \circ c)'(0) \neq 0$, where $a_i := a \circ \phi_i^{-1} \in \mathcal{A}$. Hence, by applying the weighted Nachbin result in Theorem 3.40 to the function $f \circ \phi_i^{-1} \in \mathcal{L}\mathrm{ip}^k_{\psi_i}(\phi_i(U_i))$ with curve $c_i \in C^\infty(\mathbb{R}; \phi_i(U_i))$, there exists some $a_i := a \circ \phi_i^{-1} \in \mathcal{A}_i$, with $a \in \mathcal{A}$, such that

$$\max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{\left|((f-a)\circ c)^{(j)}(t)\right|}{\psi(c(t))} = \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{\left|((f\circ\phi_i^{-1} - a_i)\circ c_i)^{(j)}(t)\right|}{\psi_i(c_i(t))} < \varepsilon,$$

Since $\varepsilon > 0$, $c \in C^\infty(\mathbb{R}; U_i)$ and $f \in \mathcal{L}\mathrm{ip}^k_\psi(X)$ were chosen arbitrarily, $\mathcal{A}$ is τ^k_ψ-dense in $\mathcal{L}\mathrm{ip}^k_\psi(X)$. □

For the vector-valued setting, we introduce polynomial algebras and define for convenient vector spaces (E, τ_E) and (F, τ_F) the set $P_n(E;F) \subset C^0(E;F)$ as the vector subspace generated by maps of the form $x \mapsto l(x)^n y$, for $n \in \mathbb{N}$, $l \in E'$ and $y \in F$. The elements of $P_n(E;F)$ are called the *n-homogeneous continuous polynomials of finite type* and $P_f(E;F)$ denotes the vector subspace generated by all $P_n(E;F)$, for $n \in \mathbb{N}$.

Definition 3.44. *$\mathcal{A} \subset C^0(X;Y)$ is called a* polynomial algebra *if $r \circ a \in \mathcal{A}$, for all $a \in \mathcal{A}$ and $r \in P_f(Y;Y)$.*

Equivalent characterizations of polynomial algebras are given in [Pro77, Lemma 4.6], e.g. $\mathcal{A} \subset C^0(X;Y)$ is a polynomial algebra if and only if $\mathcal{A}' := \{l \circ a;\, l \in Y',\, a \in \mathcal{A}\}$ is a subalgebra of $C^0(X)$ and $\mathcal{A}' \otimes Y \subset \mathcal{A}$.

Besides using polynomial algebras for the vector-valued generalization of the weighted Nachbin result in Theorem 3.43, we need to show that finite rank operators in $\mathcal{L}\mathrm{ip}^k_\psi(X) \otimes Y$ are τ^k_ψ-dense in $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$.

Definition 3.45. *(E, τ_E) has the* c^∞-approximation property *if $E' \otimes E$ is τ_{c^∞}-dense in $L(E;E)$.*

Remark 3.46. *A weaker condition is the approximation property of (E, τ_E), i.e. $E' \otimes E$ is dense in $L(E; E)$, see [Ada95]. If $C^\infty(E; E)$ is a Fréchet space, then the c^∞-topology of $C^\infty(E; E)$ is by [KM97, Lemma I.4.11] equivalent to the classical topology.*

Lemma 3.47. *Assume that (X, ψ) is a weighted space, which is a C^∞-manifold over finite dimensional vector spaces $(E_i, \tau_{E_i})_{i\in I}$. If (Y, τ_Y) admits a predual (F, τ_F) having the c^∞-approximation property, i.e. $F' \cong Y$, then $\mathcal{L}ip^k_\psi(X) \otimes Y$ is τ^k_ψ-dense in $\mathcal{L}ip^k_\psi(X; Y)$.*

Proof. We first show that if $U \subset E$ is a an open subset of a finite dimensional space (E, τ_E) and (Y, τ_Y) is as above, then $\mathcal{L}\mathrm{ip}^k(U) \otimes Y$ is c^∞-dense in $\mathcal{L}\mathrm{ip}^k(U; Y)$. By the exponential law in [FK88, Proposition 4.4.5], we observe that $\mathcal{L}\mathrm{ip}^k(U; Y) \cong \mathcal{L}\mathrm{ip}^k(U; F') \cong L(F; \mathcal{L}\mathrm{ip}^k(U))$. Since $\mathcal{L}\mathrm{ip}^k(U)$ is a convenient vector space, [KM97, Lemma I.6.9] shows that $\mathcal{L}\mathrm{ip}^k(U) \otimes Y \cong F' \otimes \mathcal{L}\mathrm{ip}^k(U)$ is c^∞-dense in $L(F; \mathcal{L}\mathrm{ip}^k(U)) \cong \mathcal{L}\mathrm{ip}^k(U; Y)$.

For the general case, it is enough to prove that $\mathcal{L}\mathrm{ip}^k_\psi(U_i) \otimes Y$ is τ^k_ψ-dense in $\mathcal{L}\mathrm{ip}^k_\psi(U_i; Y)$ since τ^k_ψ can be generated by curves $c \in C^\infty(\mathbb{R}; U_i)$, with $i \in I$. Then, it follows by the first step and because $\phi_i : U_i \to \phi_i(U_i)$ is a diffeomorphism that $\mathcal{L}\mathrm{ip}^k(U_i) \otimes Y$ is τ^k_c-dense in $\mathcal{L}\mathrm{ip}^k(U_i; Y)$, which implies that $\mathcal{L}\mathrm{ip}^k_b(U_i) \otimes Y$ is τ^k_u-dense in $\mathcal{L}\mathrm{ip}^k_b(U_i; Y)$, as τ^k_u is a weaker topology. Since τ^k_ψ is again a weaker topology on $\mathcal{L}\mathrm{ip}^k_b(U_i; Y)$ than τ^k_u, we conclude that $\mathcal{L}\mathrm{ip}^k_\psi(U_i) \otimes Y$ is τ^k_ψ-dense in $\mathcal{L}\mathrm{ip}^k_\psi(U_i; Y)$. □

Definition 3.48. *A polynomial algebra $\mathcal{A} \subset \mathcal{L}ip^k_\psi(X; Y)$ is called* locally point separating of order k *if the subalgebra $\mathcal{A}' := \{l \circ a;\ l \in Y',\ a \in \mathcal{A}\}$ is locally point separating of order k.*

Theorem 3.49 (Nachbin on $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$)**.** *Let (X,ψ) be a weighted space, which is a C^∞-manifold over finite dimensional spaces $(E_i,\tau_{E_i})_{i\in I}$, and assume that (Y,τ_Y) is a convenient vector space with predual (F,τ_F) having the c^∞-approximation property. For a polynomial algebra $\mathcal{A}\subset\mathcal{L}ip^k_\psi(X;Y)$, the following are equivalent.*

(i) $\mathcal{A}$ is locally point separating of order k, vanishes nowhere, and for every $c\in C^\infty(\mathbb{R};X)$ with $c'(0)\neq 0$ there exists some $a\in\mathcal{A}$ such that $(a\circ c)'(0)\neq 0$.

(ii) $\mathcal{A}$ is τ^k_ψ-dense in $\mathcal{L}ip^k_\psi(X;Y)$.

Proof. In order to show that (ii) implies (i), we assume by contradiction that $\mathcal{A}$ is not locally point separating of order k, or does not vanish nowhere, or there exists some $c\in C^\infty(\mathbb{R};X)$ with $c'(0)\neq 0$ such that $(a\circ c)'(0)=0$, for all $a\in\mathcal{A}$. If $\mathcal{A}$ is not locally point separating of order k, then $\mathcal{A}':=\{l\circ a;\, l\in Y',\, a\in\mathcal{A}\}$ is either not point separating at all, or there exists a point separating vector subspace $\widetilde{\mathcal{A}}'\subset\mathcal{A}'$ such that either (3.4.3) is violated, for some $\widetilde{a}'\in\widetilde{\mathcal{A}}'$, or for every $m\in\mathbb{N}_{\geqslant 1}$ and $\widetilde{\eta}'_1,...,\widetilde{\eta}'_m\in\widetilde{\mathcal{A}}'$, the function $X\ni x\mapsto\widetilde{\eta}'(x):=(\widetilde{\eta}'_1(x),...,\widetilde{\eta}'_m(x))^\top\in\mathbb{R}^m$ is not an embedding. In the first case, we define the proper closed polynomial subalgebra $\mathcal{A}_1=\big\{f\in\mathcal{L}\mathrm{ip}^k_\psi(X;Y);\, f(x_1)=f(x_2)\big\}$ of $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$, whereas in the second case, we obtain a contradiction because the vector subspace $\widetilde{\mathcal{A}}'\subset\mathcal{A}'$ contains the zero function satisfying (3.4.3). In the last case, every $U\ni x\mapsto\widetilde{\eta}'(x):=(\widetilde{\eta}'_1(x),...,\widetilde{\eta}'_m(x))^\top\in\mathbb{R}^m$ can only violate the injectivity such that there exist distinct points $x_{\widetilde{\eta}',1}\neq x_{\widetilde{\eta}',2}$ depending on $\widetilde{\eta}':U\to\mathbb{R}^m$ with $\widetilde{\eta}'(x_{\widetilde{\eta}',1})=\widetilde{\eta}'(x_{\widetilde{\eta}',2})$. From this, we define for every function $U\ni x\mapsto\widetilde{\eta}'(x):=(\widetilde{\eta}'_1(x),...,\widetilde{\eta}'_m(x))^\top\in\mathbb{R}^m$ the proper closed polynomial subalgebras $\mathcal{A}_{\widetilde{\eta}'}=\big\{f\in\mathcal{L}\mathrm{ip}^k_\psi(X;Y);\, f(x_{\widetilde{\eta}',1})=f(x_{\widetilde{\eta}',2})\big\}$ of $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$. On the other hand, if $\mathcal{A}$ does not vanish nowhere, there exists some $x_0\in X$ with $a(x_0)=0$, for all $a\in\mathcal{A}$, from which we define the proper closed subalgebra $\mathcal{A}_2=\big\{f\in\mathcal{L}\mathrm{ip}^k_\psi(U);\, f(x_0)=0\big\}$. In the third case, we define the proper closed subalgebra $\mathcal{A}_3=\big\{f\in\mathcal{L}\mathrm{ip}^k_\psi(U);\,(f\circ c)'(0)=0\big\}$. Then, the τ^k_ψ-closure of $\mathcal{A}$ is contained in one of the subalgebras $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ or $\mathcal{A}_{\widetilde{\eta}'}$, for $\widetilde{\eta}'_1,...,\widetilde{\eta}'_m\in\mathcal{A}$, such that the τ^k_ψ-closure of $\mathcal{A}$ is also a proper closed subalgebra of $\mathcal{L}\mathrm{ip}^k_\psi(U)$, which contradicts the assumption in (ii) such that the assertions in (i) must be satisfied.

For the other direction, let $f\in\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$ be a function, fix some $\varepsilon>0$, $p\in\mathcal{P}$ and $c\in C^\infty(\mathbb{R};X)$. By Lemma 3.47, there exists a function $g\in\mathcal{L}\mathrm{ip}^k_\psi(X)\otimes Y$ of the form $g(x)=\sum_{i=1}^n b_i(x)y_i$, with some $n\in\mathbb{N}_{\geqslant 1}$, $b_i\in\mathcal{L}\mathrm{ip}^k_\psi(X)$ and $y_i\in Y$,

such that

$$\max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{p\left(((f-g)\circ c)^{(j)}(t)\right)}{\psi(c(t))} < \frac{\varepsilon}{2}. \tag{3.4.12}$$

Then, the subalgebra $\mathcal{A}' = \{l \circ a;\, l \in Y',\, a \in \mathcal{A}\}$ of $\mathcal{L}\mathrm{ip}^k_\psi(X)$ satisfies the assumption of the weighted Nachbin result in Theorem 3.43 such that for every $b_i \in \mathcal{L}\mathrm{ip}^k_\psi(X)$ with $p(y_i) > 0$, there exists some $a_i \in \mathcal{A}'$ with

$$\max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{\left|((b_i-a_i)\circ c)^{(j)}(t)\right|}{\psi(c(t))} < \frac{\varepsilon}{2np(y_i)}, \tag{3.4.13}$$

whereas otherwise, if $p(y_i) = 0$, we set $a_i(x) = 0$, for all $x \in X$. By combining (3.4.12) with (3.4.13), we conclude for the function $x \mapsto a(x) := \sum_{i=1}^n a_i(x)y_i \in \mathcal{A}' \otimes Y$, satisfying $a \in \mathcal{A}$ as $\mathcal{A}' \otimes Y \subset \mathcal{A}$, that

$$\begin{aligned}\max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{p\left(((f-a)\circ c)^{(j)}(t)\right)}{\psi(c(t))} &\leqslant \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{p\left(((f-g)\circ c)^{(j)}(t)\right)}{\psi(c(t))} \\ &+ \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{p\left(((g-a)\circ c)^{(j)}(t)\right)}{\psi(c(t))} \\ &< \frac{\varepsilon}{2} + \sum_{i=1}^n p(y_i) \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{\left|((b_i-a_i)\circ c)^{(j)}(t)\right|}{\psi(c(t))} < \varepsilon.\end{aligned}$$

As $\varepsilon > 0$, $p \in \mathcal{P}$, $c \in C^\infty(\mathbb{R}; X)$ and $f \in \mathcal{L}\mathrm{ip}^k_\psi(X;Y)$ were chosen arbitrarily, $\mathcal{A}$ is τ^k_ψ-dense in $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$. □

For the infinite dimensional generalization of Theorem 3.43 and Theorem 3.49, we would additionally need to assume that (X,ψ) is a smoothly Hausdorff C^∞-manifold over convenient vector spaces $(E_i,\tau_{E_i})_{i\in I}$ having the c^∞-approximation property. Moreover, the (polynomial) algebra $\mathcal{A}$ must satisfy an additional condition such that every composition with a finite rank operator $\pi_i \in E_i' \otimes E_i$ on each model spaces (E_i,τ_{E_i}) belongs to the τ^k_ψ-closure of $\mathcal{A}$. The same idea was used in [PG76, Theorem 3.3] and [AP80, Theorem 4.3] for the approximation of differentiable functions defined on an open subset of a Banach space with respect to the topology of compact convergence and the compact-open topology, respectively.

3.4.3 Universal Approximation of Smooth Abstract Neural Networks

In this section, we prove the universal approximation property of smooth abstract neural network in this weighted setting, which generalizes the universal approximation results in [HSW90] and [Hor91] for smooth neural networks between Euclidean spaces. Hereby, we assume that the input space is a weighted space (X, ψ), which is a C^∞-manifold over finite dimensional spaces $(E_i, \tau_{E_i})_{i\in I}$. In addition, the hidden layer space (Z, ψ_Z) is given by a weighted space, which is a convenient vector space (Z, τ_Z), and the output space is also a convenient vector space (Y, τ_Y), whose topology τ_Y is generated by the family of seminorms $\mathcal{P}$, and admits a predual (F, τ_F) having the c^∞-approximation property such that $F' \cong Y$.

In this setting, a smooth abstract neural network is defined analogously as in Definition 3.15, where now all maps need to be smooth enough. For an additive family $\mathcal{H} \subset \mathcal{L}\mathrm{ip}^k_\psi(X; Z)$ consisting of hidden layer maps $h : X \to Z$, a subset $\mathcal{R} \subset \mathcal{L}\mathrm{ip}^k_{\psi_Z}(Z; Z)$ containing the activation functions, and the linear readout $\mathcal{L} \subset L(Z; Y)$, a map $\varphi : X \to Y$ is called a *smooth abstract neural network (smooth ANN)* if it is of the form

$$\varphi(x) = \sum_{k=1}^{N} \beta_k \rho_k(h_k(x)), \tag{3.4.14}$$

for $x \in X$, where $N \in \mathbb{N}_{\geqslant 1}$, $\beta_k \in \mathcal{L}$, $\rho_k \in \mathcal{R}$ and $h_k \in \mathcal{H}$. Then, the derivatives of $\varphi : X \to Y$ can be computed with Theorem 1.32(iv) and we denote by $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$ the set of smooth ANNs of the form (3.4.14).

Remark 3.50. *In [HSW90] and [Hor91], smooth neural networks were introduced between $X = \mathbb{R}^n$ and $Y = \mathbb{R}$ with hidden layer space $Z = \mathbb{R}$, where the additive family $\mathcal{H}$ consists of affine maps $h(x) = a^\top x + b$, with $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the activation family $\mathcal{R} = \{\rho\}$ is a singleton $\rho \in C^k(\mathbb{R})$, and the linear readout is given by $\mathcal{L} = \{z \mapsto wz;\, w \in \mathbb{R}\}$. Then, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}}$ is dense in $C^k(\mathbb{R}^n)$ on compacta if $\rho \in C^k(\mathbb{R})$ is either*

(i) m-finite, i.e. $0 < \int_{\mathbb{R}} \left|\rho^{(m)}(z)\right| dz < \infty$, for some $m \geqslant k$ with $\rho \in C^m(\mathbb{R})$ (see [HSW90, Corollary 3.4]),

(ii) bounded and non-constant (see [Hor91, Theorem 3+5]), or

(iii) discriminatory in the sense of Definition 2.2 (see Theorem 2.6).

If we assume $\rho \in \mathcal{L}ip^k_b(\mathbb{R})$ in addition, then $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}}$ is also τ^k_u-dense in $\mathcal{L}ip^k_b(\mathbb{R}^n)$ because τ^k_u is a weaker topology on $\mathcal{L}ip^k_b(\mathbb{R}^n)$ than τ^k_c. By using that τ^k_ψ is again a weaker topology on $\mathcal{L}ip^k_b(\mathbb{R}^n)$, where $\mathcal{L}ip^k_b(\mathbb{R}^n)$ is by definition $\|\cdot\|_{\mathcal{L}ip^k_\psi(\mathbb{R}^n)}$-dense in $\mathcal{L}ip^k_\psi(\mathbb{R}^n)$, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{\mathbb{R}^n,\mathbb{R}}$ is τ^k_ψ-dense in $\mathcal{L}ip^k_\psi(\mathbb{R}^n)$.

For the generalization of smooth neural networks to this weighted and vector-valued setting, we could consider the same abstract neural networks as in Example 3.17-3.19. Moreover, deep smooth abstract neural networks with finitely many hidden layers can be constructed by concatenation as described in Remark 3.20.

Definition 3.51. *An additive family $\mathcal{H} \subset \mathcal{L}ip^k_\psi(X;Z)$ between X and Z is called* strong *if*

(i) for every $c \in C^\infty(\mathbb{R};X)$ with $c'(0) \neq 0$ there exists some $h \in \mathcal{H}$ such that $(h \circ c)'(0) \neq 0$, and

(ii) for every $i \in I$ there exists some $m \in \mathbb{N}_{\geqslant 1}$ and $h_1, ..., h_m \in \mathcal{H}$ as well as $z'_1, ..., z'_m \in Z'$ such that $U_i \ni x \mapsto (z'_1(h_1(x)), ..., z'_m(h_m(x)))^\top \in \mathbb{R}^m$ is an embedding, i.e. continuous and injective.

Theorem 3.52 (Universal Approximation on $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$)**.** *For a strong additive family $\mathcal{H} \subset \mathcal{L}ip^k_\psi(X;Z)$, assume that the architecture $(\mathcal{H},\mathcal{R},\mathcal{L})$ between X and Y be compatible with another architecture $(\mathcal{G},\mathcal{R},\mathcal{K})$ between Z and Z. If $Z' \subset \mathcal{L}ip^k_{\psi_Z}(Z)$ and $\mathcal{NN}^{\mathcal{G},\mathcal{R},\mathcal{K}}_{Z,Z}$ is $\tau^k_{\psi_Z}$-dense in $\mathcal{L}ip^k_{\psi_Z}(Z;Z)$, then $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$ is τ^k_ψ-dense in $\mathcal{L}ip^k_\psi(X;Y)$, i.e. for every $f \in \mathcal{L}ip^k_\psi(X;Y)$, $\varepsilon > 0$ and $q \in \mathcal{Q}$ there exists $\varphi \in \mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$ such that*

$$q(f-\varphi) = \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{p\left(((f-\varphi)\circ c)^{(j)}(t)\right)}{\psi(c(t))} < \varepsilon,$$

where $p \in \mathcal{P}$ and $c \in C^\infty(\mathbb{R};X)$ are the seminorm and the curve correponding to $q \in \mathcal{Q}$.

Proof. Let $\mathcal{A} = \operatorname{span}\{x \mapsto z'(h(x))^n y_0;\ n\in\mathbb{N},\ z'\in Z',\ h\in\mathcal{H},\ y_0\in Y\}$ be a polynomial algebra, which is contained in $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$, because it holds by assumption that $Z' \subset \mathcal{L}\mathrm{ip}^k_{\psi_Z}(Z)$.

First, we show that $\mathcal{A}$ lies again in the τ^k_ψ-closure of $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$. Let $\mathcal{P}_Z$ be a family of seminorms generating the topology τ_Z of Z, which exists by Theorem 1.25, fix some $\varepsilon > 0$, $p \in \mathcal{P}$, $c \in C^\infty(\mathbb{R};X)$ and consider the function $x \mapsto a(x) := z'(h(x))^n y_0 \in \mathcal{A}$, for some $n\in\mathbb{N}$, $z'\in Z'$, $h\in\mathcal{H}$ and $y_0 \in Y$, where we define $C_1 := \sup_{x\in X} \psi_Z(h(x))/\psi(x)$. Then, there exists some $\gamma \in \Gamma := \{\gamma \in L(Z;Y);\ \forall \alpha\in\mathcal{K} : \gamma\circ\alpha\in\mathcal{L}\}$ and $z_0 \in Z$ such that $p(y_0 - \gamma(z_0)) < \varepsilon/(2C_2)$, where $C_2 := \max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \left|(a\circ c)^{(j)}(t)\right|/\psi(c(t))$. Moreover, we define $C_3 := \sup_{z\in Z,\, p_Z(z)\leqslant 1} p(\gamma(z))$ and $c_Z := h\circ c \in \mathcal{L}\mathrm{ip}^k(\mathbb{R};Z)$. As $\mathcal{NN}^{\mathcal{G},\mathcal{R},\mathcal{K}}_{Z,Z}$ is $\tau^k_{\psi_Z}$-dense in $\mathcal{L}\mathrm{ip}^k_{\psi_Z}(Z;Z)$, there exists for the function $z \mapsto a_Z(z) := z'(z)^n z_0 \in \mathcal{L}\mathrm{ip}^k_{\psi_Z}(Z;Z)$ some $\varphi_Z \in \mathcal{NN}^{\mathcal{G},\mathcal{R},\mathcal{K}}_{Z,Z}$ with

$$\max_{j=0,\dots,k} \sup_{t\in\mathbb{R}} \frac{p_Z\left(((a_Z-\varphi_Z)\circ c_Z)^{(j)}(t)\right)}{\psi_Z(c_Z(t))} < \frac{\varepsilon}{2C_1C_3}, \tag{3.4.15}$$

where we have used that the topology $\tau^k_{\psi_Z}$ on $\mathcal{L}\mathrm{ip}^k_{\psi_Z}(Z;Z)$ can equivalently be generated by $\mathcal{L}\mathrm{ip}^k$-curves instead of smooth curves, see [KM97, Theorem III.12.11]. For the function $x \mapsto b(x) := z'(h(x))^n\gamma(z_0)$, we define $\varphi := \gamma\circ\varphi_Z\circ h \in \mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$, by using that $g\circ h\in\mathcal{H}$, for all $g\in\mathcal{G}$, such that (3.4.15)

implies that

$$\max_{j=0,...,k} \sup_{t\in\mathbb{R}} \frac{p\left(((a-\varphi)\circ c)^{(j)}(t)\right)}{\psi(c(t))} \leqslant C_2 p(y_0-\gamma(z_0))$$
$$+ \max_{j=0,...,k} \sup_{t\in\mathbb{R}} \frac{p\left(((b-\varphi)\circ c)^{(j)}(t)\right)}{\psi(c(t))} < C_2\frac{\varepsilon}{2C_2}$$
$$+ C_1C_3 \max_{j=0,...,k} \sup_{t\in\mathbb{R}} \frac{p_Z\left(((a_Z-\varphi_Z)\circ c_Z)^{(j)}(t)\right)}{\psi_Z(c(t))} < \varepsilon.$$

Since $\varepsilon > 0$ and $p \in \mathcal{P}$ were chosen arbitrarily, it follows that $x \mapsto a(x) = z'(h(x))^n y_0 \in \mathcal{A}$ belongs to the τ_ψ^k-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$. By the triangle inequality, also linear combinations of such functions are contained in the τ_ψ^k-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ such that the entire polynomial algebra $\mathcal{A}$ lies in the τ_ψ^k-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$.

Finally, we apply the weighted Nachbin result in Theorem 3.49 to show that $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is τ_ψ^k-dense in $\mathcal{L}\mathrm{ip}_\psi^k(X;Y)$. For this, we observe that $\mathcal{A}$ is locally point separating of order k with point separating vector subspace $\widetilde{\mathcal{A}}' := \{z'\circ h;\, z'\in Z',\, h\in\mathcal{H}\}$, where we have used that $\mathcal{H}$ and Z' are point separating, and where we have employed that $\mathcal{H}$ is strong such there exists for every $i \in I$ some $m \in \mathbb{N}_{\geqslant 1}$, $h_1, ..., h_m \in \mathcal{H}$ and $z'_1, ..., z'_m \in Z'$ such that $U_i \ni x \mapsto \widetilde{\eta}'(x) := (\widetilde{\eta}_1(x), ..., \widetilde{\eta}_m(x))^\top \in \mathbb{R}^m$ is an embedding, with $\widetilde{\eta}_j := z'_j \circ h_j \in \widetilde{\mathcal{A}}'$, for $j = 1, ..., m$. Moreover, $\mathcal{A}$ vanishes nowhere, because for every $x_0 \in X$ there exist some $h \in \mathcal{H}$ and $z' \in Z'$ such that $x \mapsto a(x) := z'(h(x))y_0 \in \mathcal{A}$, with some $y_0 \in Y\backslash\{0\}$, satisfies $a(x_0) \neq 0$. Furthermore, there exists for every $c \in C^\infty(\mathbb{R}; X)$ with $c'(0) \neq 0$ by using that $\mathcal{H}$ is strong some $h \in \mathcal{H}$ and $z' \in Z'$, with $z'((h\circ c)'(0)) \neq 0$, such that $x \mapsto a(x) := z'(h(x))y_0 \in \mathcal{A}$, with some $y_0 \in Y\backslash\{0\}$, satisfies $(a\circ c)'(0) = z'((h\circ c)'(0))\, y_0 \neq 0$. Hence, by applying the weighted Nachbin result in Theorem 3.49, it follows that $\mathcal{A}$ is τ_ψ^k-dense in $\mathcal{L}\mathrm{ip}_\psi^k(X;Y)$. However, because $\mathcal{A}$ lies in the τ_ψ^k-closure of $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$, we conclude that $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is also τ_ψ^k-dense in $\mathcal{L}\mathrm{ip}_\psi^k(X;Y)$. □

Theorem 3.52 is a generalization of [HSW90, Theorem 3.1] to this weighted setting, where we obtain universal approximation on a compact subset $K \subset X$ by choosing $\psi(x) = 1$ and applying Theorem 3.52. Moreover, the argument in Theorem 3.52 can be iterated to obtain universal approximation of deep smooth abstract neural network with finitely many hidden layers. In addition, it is possible to consider an infinite dimensional hidden layer space (Z, ψ_Z) and several activation functions in $\mathcal{R} \subset \mathcal{L}\mathrm{ip}_{\psi_Z}^k(Z;Z)$ such that $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is

under some additional assumptions τ_ψ^k-dense in $\mathcal{L}\mathrm{ip}_\psi^k(X;Y)$ if for every hidden layer map $h \in \mathcal{H}$ there exists a direction given by $z_0 \in Z\backslash\{0\}$, with $x \mapsto j(h(x))z_0 \in \mathcal{H}$, for some $j \in Z'$, and there exist $\rho \in \mathcal{R}$ and $z' \in Z'$ such that $\mathbb{R} \ni t \mapsto z'(\rho(tz_0)) \in \mathbb{R}$ satisfies one of the conditions in Remark 3.50.

3.5 Randomized Abstract Neural Networks

Abstract neural networks can be used to learn a function between infinite dimensional spaces. However, the training of the parameters $h_k \in \mathcal{H}$ and $\beta_k \in \mathcal{L}$ is a non-linear regression task, which means a general non-convex optimization problem. In addition, by taking into account infinite dimensional objects such as paths or functions, the number of parameters in the abstract neural network can become very large to ensure satisfactory approximation, which in turn leads to an unstable or unfeasible training procedure.

To overcome this issue, we borrow ideas from the field of reservoir computing, where the key idea is that the inner parameters of the abstract neural network are initialized randomly and only the linear readout is trained. This brings us back to a convex optimization problem (if the loss function is also convex) and by choosing the size of the hidden layer large enough, we obtain a sufficiently rich random basis of the hidden layer. In the literature, [RR07] and [GGO20] showed that this approach is successful for conventional neural networks between finite dimensional Euclidean spaces, whereas [HTW19] proved that the trained randomized neural network behaves asymptotically like the spline interpolation. Moreover, it was shown in [KB20, Section 3.3] that training only the linear readout can be generalized to infinite dimensional input and output spaces, which was also applied to echo-state networks (ESNs) as mentioned in Section 2.4.

In our infinite dimensional setting, we assume that the input space (X, ψ) is a separable metrizable weighted space, the hidden layer space (Z, ψ_Z) is a weighted Banach space with norm $\|\cdot\|_Z$, and the output space $(Y, \|\cdot\|_Y)$ is another Banach space. Furthermore, we suppose that $\mathcal{H} \subset \mathcal{B}_\psi(X;Y)$ is an additive family between X and Z, $\mathcal{L} \subset L(Z;Y)$ is a linear readout between Z and Y, and $\mathcal{R} \subset \mathcal{B}_{\psi_Z}(Z;Z)$ contains the activation functions on Z such that $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(X;Y)$, which means that the abstract neural networks in $\mathcal{NN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ admit the universal approximation property.

In order to include randomness, we assume that there exists a probability space $(\Omega, \Sigma, \mathbb{P})$ with locally compact topological space (Ω, τ_Ω) such that $h : \Omega \to \mathcal{H}$ defines an $\mathcal{H}$-valued random variable, where we suppose that the distribution

of $h : \Omega \to \mathcal{H}$ is such that for every distinct pair $x_1, x_2 \in X$, we have

$$\begin{aligned} \mathbb{P}\left[\{\omega \in \Omega;\, h(\omega)(x_1) \neq h(\omega)(x_2)\}\right] > 0, \\ \mathbb{P}\left[\{\omega \in \Omega;\, h(\omega)(x_1) \neq 0\}\right] > 0. \end{aligned} \tag{3.5.1}$$

Let $(h_k)_{k\in\mathbb{N}} \sim h$ be a sequence of independent and identically distributed (i.i.d.) random variables, from which we define for some $N \in \mathbb{N}_{\geqslant 1}$ a *randomized abstract neural network (RANN)* $\varphi : \Omega \times X \to Y$ by

$$\varphi(\omega)(x) = \sum_{k=1}^{N} \beta_k \rho_k \left(h_k(\omega)(x)\right), \tag{3.5.2}$$

for $\omega \in \Omega$ and $x \in X$, where $\rho_k \in \mathcal{R}$ and $\beta_k \in \mathcal{L}$. Hereby, we denote by $\mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ the set of such randomized abstract neural networks in (3.5.2) and by initializing the hidden layer maps $h_1, ..., h_N \in \mathcal{H}$ randomly, it only remains to train the linear readouts $\beta_1, ..., \beta_N \in \mathcal{L}$ given the realizations $h_1(\omega), ..., h_N(\omega) \in \mathcal{H}$. The following lemma ensures in particular that $\varphi : \Omega \times X \to Y$ is a random function.

Lemma 3.53. *The randomized abstract neural network $\varphi : \Omega \times X \to Y$ in (3.5.2) is product-measurable, i.e. $\Sigma \otimes \mathcal{F}/\mathcal{T}$-measurable, where $\mathcal{F}$ and $\mathcal{T}$ denote the Borel σ-algebra of X and Y, respectively.*

Proof. First, we observe that $\omega \mapsto h_k(\omega)(x)$ is $\Sigma/\mathcal{S}$-measurable, for all $k \in \mathbb{N}$ and $x \in X$, where $\mathcal{S}$ denotes the Borel σ-algebra of Z. By continuity of $z \mapsto \sum_{k=1}^{N} \beta_k \rho_k(z)$ on every compact preimage $K_{Z,R} := \psi_Z^{-1}(]0, R])$, we conclude that $\omega \mapsto \varphi(\omega)(x)$ is $\Sigma/\mathcal{T}$-measurable, for all $x \in X$. On the other hand, we observe that $x \mapsto \varphi(\omega)(x)$ is by definition continuous on every compact preimage $K_R = \psi^{-1}(]0, R])$, for all $\omega \in \Omega$ and $R > 0$, such that $(\omega, x) \mapsto \varphi(\omega)(x)$ is by [AB06, Lemma 4.51] $\Sigma \otimes \mathcal{F}_R/\mathcal{T}$-measurable, where $\mathcal{F}_R$ denotes the Borel σ-algebra of K_R. Because this holds for every $R > 0$, $(\omega, x) \mapsto \varphi(\omega)(x)$ is $\Sigma \otimes \mathcal{F}/\mathcal{T}$-measurable. □

Now, we consider a randomized ANN $\varphi : \Omega \to \mathcal{B}_\psi(X; Y)$ as random variable in $L^p(\Omega; \mathcal{B}_\psi(X; Y))$, with $1 \leqslant p < \infty$, where we assume that the random variable $\rho_k \circ h : \Omega \to \mathcal{B}_\psi(X; Z)$ belongs to $L^p(\Omega; \mathcal{B}_\psi(X; Z))$ such that $\mathbb{E}\left[\|\rho \circ h\|^p_{\mathcal{B}_\psi(X;Z)}\right] < \infty$, for all $\rho \in \mathcal{R}$, which implies by Minkowski's inequality that

$$\|\varphi\|_{L^p(\Omega;\mathcal{B}_\psi(X;Y))} := \mathbb{E}\left[\|\varphi\|^p_{\mathcal{B}_\psi(X;Y)}\right]^{\frac{1}{p}} \leqslant \sum_{k=1}^{N} \|\beta_k\|_{L(Z;Y)} \mathbb{E}\left[\|\rho_k \circ h\|^p_{\mathcal{B}_\psi(X;Z)}\right]^{\frac{1}{p}}$$

is finite, such that $\varphi \in L^p(\Omega; \mathcal{B}_\psi(X;Y))$ is well-defined, where we have used that h_1, h_2, ... are identically distributed. In the following, we show that any random function $f \in L^p(\Omega; \mathcal{B}_\psi(X;Y))$ and therefore any deterministic function $f \in \mathcal{B}_\psi(X;Y)$ can be approximated by a randomized abstract neural network $\varphi \in \mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$.

Example 3.54. *For $X = \mathbb{R}^n$ and $Z = Y = \mathbb{R}$ with weights $\psi(x) = \exp\left(\|x\|^2\right)$ and $\psi_Z(z) = \exp(|z|)$, we consider $\mathcal{H} = \left\{x \mapsto a^\top x + b;\, a \in \mathbb{R}^n,\, b \in \mathbb{R}\right\}$, an activation function $\rho \in C_b^0(\mathbb{R})$ and a linear readout $\mathcal{L} = \{z \mapsto wz;\, w \in \mathbb{R}\}$ such that $\mathcal{NN}_{\mathbb{R}^n,\mathbb{R}}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(\mathbb{R}^n)$. Then, we define $h : \Omega \to \mathcal{H}$ as $h(\omega) = (x \mapsto \sum_{i=1}^n a_i(\omega)x_i + b(\omega))$, for independent random variables $a_i, b \sim \mathcal{N}(0,1)$ following a normal distribution. Because $\rho : \mathbb{R} \to \mathbb{R}$ is bounded, we have $\mathbb{E}\left[\|\rho \circ h\|_{\mathcal{B}_\psi(\mathbb{R}^n)}^p\right] < \infty$, for all $1 \leqslant p < \infty$.*

Lemma 3.55 ([HNVW16, Lemma 1.2.31])**.** *If (Ω, τ_Ω) is locally compact, then the space $C_c^0(\Omega; \mathcal{B}_\psi(X;Y))$ of compactly supported random variables is dense in $L^p(\Omega; \mathcal{B}_\psi(X;Y))$ with respect to $\|\cdot\|_{L^p(\Omega;\mathcal{B}_\psi(X;Y))}$.*

Theorem 3.56 (Random Universal Approximation)**.** *Let $1 \leqslant p < \infty$ and assume that the architecture $(\mathcal{H},\mathcal{R},\mathcal{L})$ between X and Y is compatible with another architecture $(\mathcal{G},\mathcal{R},\mathcal{K})$ between Z and Z such that $\mathcal{NN}_{Z,Z}^{\mathcal{G},\mathcal{R},\mathcal{K}}$ is $\tau_{\psi_Z}^0$-dense in $\mathcal{B}_{\psi_Z}(Z;Z)$. If the random variable $h : \Omega \to \mathcal{H}$ with $h \in L^p(\Omega; \mathcal{B}_\psi(X;Z))$ satisfies (3.5.1), then $\mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is dense in $L^p(\Omega; \mathcal{B}_\psi(X;Y))$, i.e. for every $f \in L^p(\Omega; \mathcal{B}_\psi(X;Y))$ and $\varepsilon > 0$ there exists some $\varphi \in \mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ such that*

$$\|f - \varphi\|_{L^p(\Omega;\mathcal{B}_\psi(X;Y))} = \mathbb{E}\left[\left(\sup_{x\in X} \frac{\|f(x) - \varphi(x)\|_Y}{\psi(x)}\right)^p\right]^{1/p} < \varepsilon.$$

Proof. First, we show the approximation of $f \in C_c^0(\Omega; \mathcal{B}_\psi(X;Y))$ by some $\varphi \in \mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ in a weighted supremum norm. For this purpose, let $f \in C_c^0(\Omega; \mathcal{B}_\psi(X;Y))$ with support contained in a compact set $K \subset \Omega$ and fix some $\varepsilon > 0$. Let (X_0, ψ_{X_0}) be the weighted space with $X_0 = K \times X$ and $\psi_{X_0}(\omega, x) = \psi(x)$, for all $(\omega, x) \in K \times X$, and define the additive family $\mathcal{H}_0 = \{(h_1 + ... + h_n + b)|_{X_0} : X_0 \to Z;\, n \in \mathbb{N}_{\geqslant 1},\, h_i \sim h,\, b \in Z\}$, which implies that $\mathcal{NN}_{X_0,Y}^{\mathcal{H}_0,\mathcal{R},\mathcal{L}} = \mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}\Big|_{X_0}$. In this case, the compatibility of $g \circ h \in \mathcal{H}$, for all $g \in \mathcal{G}$ and $h \in \mathcal{H}$, implies $g \circ h_0 \in \mathcal{H}$, for all $g \in \mathcal{G}$ and $h_0 \in \mathcal{H}_0$.

Moreover, for every $h_0 \in \mathcal{H}_0$ there exists by additivity of $\mathcal{H}$ some $h \in \mathcal{H}$ such that $\sup_{(\omega,x)\in X_0} \psi_Z(h_0(\omega)(x))/\psi_{X_0}(\omega,x) \leqslant \sup_{x\in X} \psi_Z(h(x))/\psi(x) < \infty$, which shows that the architecture $(\mathcal{H}_0, \mathcal{R}, \mathcal{L})$ between X_0 and Y is also compatible with the architecture $(\mathcal{G}, \mathcal{R}, \mathcal{K})$ between Z and Z. Hence, we conclude by applying Theorem 3.27 that there exists some $\varphi \in \mathcal{RN}_{X,Y}^{\mathcal{H}_0,\mathcal{R},\mathcal{L}}$ such that

$$\sup_{(\omega,x)\in\Omega\times X} \frac{\|f(\omega)(x) - \varphi(\omega)(x)\|_Y}{\psi_{X_0}(\omega,x)} = \sup_{(\omega,x)\in\Omega\times X} \frac{\|f(\omega)(x) - \varphi(\omega)(x)\|_Y}{\psi(x)} < \varepsilon,$$

which shows that $\mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is dense in $C_c^0(\Omega; \mathcal{B}_\psi(X;Y))$, because $\varepsilon > 0$ and $f \in C_c^0(\Omega; \mathcal{B}_\psi(X;Y))$ were chosen arbitrarily. Since the topology induced by $\|\cdot\|_{L^p(\Omega;\mathcal{B}_\psi(X;Y))}$ is a weaker topology on $C_c^0(\Omega; \mathcal{B}_\psi(X;Y))$ than $\tau_{\psi_{X_0}}^0$, where $C_c^0(\Omega; \mathcal{B}_\psi(X;Y))$ is by Lemma 3.55 dense in $L^p(\Omega; \mathcal{B}_\psi(X;Y))$, the set of randomized abstract neural network $\mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ is also dense in $L^p(\Omega; \mathcal{B}_\psi(X;Y))$. □

By an application of the Markov inequality, it follows from Theorem 3.56 that RANNs are able to approximate any function $f \in \mathcal{B}_\psi(X;Y)$ with an error that rarely exceeds a certain tolerance $\varepsilon > 0$.

Corollary 3.57. *Let $1 \leqslant p < \infty$ and assume that the architecture $(\mathcal{H}, \mathcal{R}, \mathcal{L})$ between X and Y is compatible with another architecture $(\mathcal{G}, \mathcal{R}, \mathcal{K})$ between Z and Z such that $\mathcal{NN}_{Z,Z}^{\mathcal{G},\mathcal{R},\mathcal{K}}$ is $\tau_{\psi_Z}^0$-dense in $\mathcal{B}_{\psi_Z}(Z;Z)$. If the random variable $h : \Omega \to \mathcal{H}$ with $h \in L^p(\Omega; \mathcal{B}_\psi(X;Z))$ satisfies (3.5.1), then for every $f \in L^p(\Omega; \mathcal{B}_\psi(X;Y))$ and $\delta, \varepsilon > 0$ there exists some $\varphi \in \mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ such that*

$$\mathbb{P}\left[\left\{\omega \in \Omega;\, \sup_{x\in X} \frac{\|f(\omega)(x) - \varphi(\omega)(x)\|_Y}{\psi(x)} \geqslant \varepsilon\right\}\right] < \delta.$$

Proof. Let $f \in L^p(\Omega; \mathcal{B}_\psi(X;Y))$ and fix $\delta, \varepsilon > 0$. Then, there exists by Theorem 3.56 some $\varphi \in \mathcal{RN}_{X,Y}^{\mathcal{H},\mathcal{R},\mathcal{L}}$ satisfying $\mathbb{E}\left[\|f - \varphi\|_{\mathcal{B}_\psi(X;Y)}^p\right] < \delta\varepsilon^p$ such that the conclusion follows by applying the Markov inequality. □

Now, we apply these arguments to the universal approximation result with derivatives in Theorem 3.52 to show that smooth RANNs are able to approximate differentiable functions arbitrarily well. For this purpose, we assume that the input space (X, ψ) is a weighted space, which is a C^∞-manifold over finite dimensional spaces $(E_i, \|\cdot\|_{E_i})_{i\in I}$, the hidden layer space (Z, ψ_Z) is a Banach space with norm $\|\cdot\|_Z$, and the output space $(Y, \|\cdot\|_Y)$ is also a Banach space with predual $(F, \|\cdot\|_F)$ having the approximation property.

Theorem 3.58 (Random Smooth Universal Approximation)**.** *Let $1 \leqslant p < \infty$ and let $\mathcal{H} \subset \mathcal{L}ip^k_\psi(X;Z)$ be a strong additive family. Assume that the architecture $(\mathcal{H},\mathcal{R},\mathcal{L})$ between X and Y is compatible with $(\mathcal{G},\mathcal{R},\mathcal{K})$ between Z and Z such that $\mathcal{NN}^{\mathcal{G},\mathcal{R},\mathcal{K}}_{Z,Z}$ is $\tau^k_{\psi_Z}$-dense in $\mathcal{L}ip^k_{\psi_Z}(Z;Z)$. If $h : \Omega \to \mathcal{H}$ with $h \in L^p(\Omega;\mathcal{L}ip^k_\psi(X;Z))$ satisfies* (3.5.1) *and*

$$\mathbb{P}\left[\left\{\omega \in \Omega;\ (h(\omega)\circ c)'(0) \neq 0\right\}\right] > 0, \tag{3.5.3}$$

for all $c \in C^\infty(\mathbb{R};X)$, then $\mathcal{RNN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$ is dense in $L^p(\Omega;\mathcal{L}ip^k_\psi(X;Y))$, i.e. for every $f \in L^p(\Omega;\mathcal{L}ip^k_\psi(X;Y))$ and $\varepsilon > 0$ as well as $c \in C^\infty(\mathbb{R};X)$ there exists some $\varphi \in \mathcal{RNN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$ such that

$$\mathbb{E}\left[\left(\max_{j=0,\dots,k}\sup_{t\in\mathbb{R}}\frac{\left\|((f-\varphi)\circ c)^{(j)}(t)\right\|_Y}{\psi(c(t))}\right)^p\right]^{1/p} < \varepsilon.$$

Proof. Since $\mathcal{L}\mathrm{ip}^k_\psi(X;Y)$ is a locally convex topological vector space, we can use the arguments of [HNVW16, Proposition 1.2.2, Lemma 1.2.19(i)] applied to every $c \in C^\infty(\mathbb{R};X)$ such that simple functions of the form $\omega \mapsto \sum_{m=1}^n \mathbb{1}_{A_m}(\omega) f_m$, for $n \in \mathbb{N}_{\geqslant 1}$, $A_m \in \Sigma$ and $f_m \in \mathcal{L}\mathrm{ip}^k_\psi(X;Y)$, are dense in $L^p(\Omega;\mathcal{L}\mathrm{ip}^k_\psi(X;Y))$. Hence, the arguments of [HNVW16, Proposition 1.2.31] are still valid such that $C^0_c(\Omega;\mathcal{L}\mathrm{ip}^k_\psi(X;Y))$ is dense in $L^p(\Omega;\mathcal{L}\mathrm{ip}^k_\psi(X;Y))$. The remaining proof follows along the lines of the proof of Theorem 3.56, where we use the universal approximation result in Theorem 3.52 instead of Theorem 3.27. □

Corollary 3.59. *Let $1 \leqslant p < \infty$ and let $\mathcal{H} \subset \mathcal{L}ip^k_\psi(X;Z)$ be a strong additive family. Assume that $(\mathcal{H},\mathcal{R},\mathcal{L})$ between X and Y is compatible with $(\mathcal{G},\mathcal{R},\mathcal{K})$ between Z and Z such that $\mathcal{NN}^{\mathcal{G},\mathcal{R},\mathcal{K}}_{Z,Z}$ is $\tau^k_{\psi_Z}$-dense in $\mathcal{L}ip^k_{\psi_Z}(Z;Z)$. If the random variable $h : \Omega \to \mathcal{H}$ with $h \in L^p(\Omega;\mathcal{L}ip^k_\psi(X;Z))$ satisfies* (3.5.1) *and* (3.5.3)*, then for every $f \in L^p(\Omega;\mathcal{L}ip^k_\psi(X;Y))$ and $\delta,\varepsilon > 0$ as well as $c \in C^\infty(\mathbb{R};X)$ there exists some $\varphi \in \mathcal{RNN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$ such that*

$$\mathbb{P}\left[\left\{\omega\in\Omega;\ \max_{j=0,\dots,k}\sup_{t\in\mathbb{R}}\frac{\left\|((f(\omega)-\varphi(\omega))\circ c)^{(j)}(t)\right\|_Y}{\psi(c(t))} \geqslant \varepsilon\right\}\right] < \delta.$$

Proof. The proof follows by applying the same arguments including Markov's inequality as in the proof of Corollary 3.57, where Theorem 3.58 instead of

Theorem 3.56 is applied. □

Moreover, the arguments in Theorem 3.56, Corollary 3.57, Theorem 3.58 and Corollary 3.59 could be generalized to locally convex topological vector spaces as hidden and output layers.

3.6 Lipschitz Functions and Signature Methods

In this section, we show another application of the weighted Stone-Weierstrass stated in Theorem 3.11, which is related to the theory of rough paths introduced by Terry Lyons, see [LCL07]. Hereby, the signature of a path consisting of its iterated integrals can be used to approximate any path space functional arbitrarily well. For example, [CO18] and [KO19] introduced kernel-based methods to learn functionals depending on the signature, whereas [CGOT20] considered a randomized version of the discrete-time signature. These novel techniques have been succesfully applied for hedging and optimal stopping problems in Finance, see for example [LNP19] and [JLP20]. Opposed to these results, we consider here the non-compact input space of Lipschitz functions equipped with a weak-$*$-topology, where we follow the notation of [Wea99].

3.6.1 Predual of Lipschitz Functions

For a pointed metric space (S, d_S), i.e. a metric space with designated origin $0 \in S$, we will show the existence of a Banach space $Æ(S)$, the so-called *Arens-Eells space*, whose dual can be naturally identified with the space of Lipschitz functions preserving the origin. More precisely, let $(Y, \|\cdot\|_Y)$ be a Banach space, for which a map $f : S \to Y$ is called *Lipschitz* if there exists a constant $C > 0$ such that

$$\|f(s) - f(t)\|_Y \leqslant C d_S(s, t), \tag{3.6.1}$$

for all $s, t \in S$. Using this, we introduce the following two spaces of Lipschitz functions.

Definition 3.60. *For a metric space (S, d_S) and a Banach space $(Y, \|\cdot\|_Y)$, we introduce the following.*

(i) The Lipschitz space $\mathrm{Lip}(S;Y)$ *is defined as bounded Lipschitz functions $f : S \to Y$ with norm $\|f\|_{\mathrm{Lip}(S;Y)} = \max\left(L(f), \sup_{s\in S} \|f(s)\|_Y\right)$, where $L(f)$ denotes the Lipschitz constant of $f : S \to Y$, which is defined as the smallest number $C > 0$ such that (3.6.1) holds true.*

(ii) If (S, d_S) is additionally pointed, we define the Lipschitz space $\mathrm{Lip}_0(S;Y)$ *as Lipschitz functions $f : S \to Y$ preserving the origin, i.e. $f(0) = 0$, which is equipped with the norm $\|f\|_{\mathrm{Lip}_0(S;Y)} = L(f)$.*

In both cases, we abbreviate $\mathrm{Lip}(S) := \mathrm{Lip}(S;\mathbb{R})$ *and* $\mathrm{Lip}_0(S) := \mathrm{Lip}_0(S;\mathbb{R})$ *if* $Y = \mathbb{R}$.

For every (pointed) metric space (S, d_S), $\mathrm{Lip}(S;Y)$ and $\mathrm{Lip}_0(S;Y)$ are again Banach spaces with their corresponding norm, see [Wea99, Proposition 2.3]. We start by showing that $\mathrm{Lip}(S)$ and $\mathrm{Lip}_0(S)$ admit a predual, with the help of the so-called *De Leeuw's map* $\Phi : \mathrm{Lip}_0(S) \to B(\widetilde{S})$, which is defined by

$$\Phi(f)(s,t) = \frac{f(s) - f(t)}{d_S(s,t)},$$

for a function $f \in \mathrm{Lip}_0(S)$ and $(s,t) \in \widetilde{S}$, where $\widetilde{S} = \{(s,t) \in S \times S;\, s \neq t\}$ denotes the space $S \times S$ without diagonal, and where $B(\widetilde{S})$ denotes the space of bounded functions $f : \widetilde{S} \to \mathbb{R}$ on $\widetilde{S}$ such that $\sup_{(s,t)\in\widetilde{S}} |f(s,t)| < \infty$, see [Wea99, Theorem 2.35].

Lemma 3.61. *For a compact metric space (S, d_S), the following holds true:*

(i) $\mathrm{Lip}(S)$ *is a dual space and on bounded subsets of* $\mathrm{Lip}(S)$*, the weak-*-topology coincides with the topology of uniform convergence.*

(ii) If (S, d_S) is additionally pointed, then $\mathrm{Lip}_0(S)$ *is a dual space and on bounded subsets of* $\mathrm{Lip}_0(S)$*, the weak-*-topology coincides with the topology of uniform convergence.*

Proof. This follows by combining [Wea99, Theorem 2.37] and [Wea99, Proposition 2.39] for (ii) as well as [Wea99, Corollary 2.38] and [Wea99, Corollary 2.40] for (i). □

Now, we introduce the vector space $\mathcal{M}$ of so-called *molecules*, which are defined as functions $m : S \to \mathbb{R}$ with finite support such that $\sum_{s\in S} m(s) = 0$.

In addition, we define the pairing $\langle f, m\rangle := \sum_{s\in S} f(s)m(s)$ such that for every $f \in \mathrm{Lip}_0(S)$ and $m \in \mathcal{M}$, it holds that

$$|\langle f, m\rangle| \leqslant \|f\|_{\mathrm{Lip}_0(S)}\|m\|_{Æ(S)},$$

where $\|\cdot\|_{Æ(S)}$ is defined in the following. An *elementary molecule* is defined as $m_{st} = \mathbb{1}_{\{s\}} - \mathbb{1}_{\{t\}} \in \mathcal{M}$, for $s, t \in S$, such that every molecule $m \in \mathcal{M}$ has an expression as linear combination of elementary molecules.

Definition 3.62. *The* Arens-Eells norm *of a molecule $m \in \mathcal{M}$ is defined as*

$$\|m\|_{Æ(S)} = \inf\left\{\sum_{i=1}^{n} |a_i| d_S(s_i, t_i);\ m = \sum_{i=1}^{n} a_i m_{s_i t_i},\ s_i, t_i \in S\right\},$$

where the infimum is taken over all finite linear combinations of elementary molecules. In this case, the Arens-Eells space $Æ(S)$ *is defined as the $\|\cdot\|_{Æ(S)}$-closure of $\mathcal{M}$, which becomes with $\|\cdot\|_{Æ(S)}$ a Banach space.*

In the non-linear Banach space literature, the Arens-Eells space $Æ(S)$ is known as the Lipschitz-free space and denoted by $\mathcal{F}(S)$, see e.g. [GK03]. Now, we can identify the predual of $\mathrm{Lip}_0(S)$ with $Æ(S)$.

Theorem 3.63. *Let (S, d) be a pointed compact metric space. Then, $Æ(S)^*$ is isometrically isomorphic to $\mathrm{Lip}_0(S)$ and on bounded subsets of $\mathrm{Lip}_0(S)$, the weak-$*$-topology coincides with the topology of uniform convergence.*

Proof. This follows by combining [Wea99, Theorem 3.3] and [Wea99, Proposition 2.39]. □

For the vector-valued case with finite dimensional $(Y, \|\cdot\|_Y)$, we use that the dual of Y^* is equal to Y, and consider the projective tensor product $Æ(S)\hat{\otimes}_\pi Y^*$, which consists by [SZ06, Theorem III§6.4] and because Y^* is finite dimensional of elements of the form $T = \sum_{j=1}^{n} m_j \otimes l_j$, for $n \in \mathbb{N}$, $m_j \in Æ(S)$ and $l_j \in Y^*$, and which is equipped with the norm $\|T\|_{Æ(S)\hat{\otimes}_\pi Y^*} = \sum_{j=1}^{m} \|m_j\|_{Æ(S)}\|l_j\|_{Y^*}$.

Theorem 3.64 ([Wea99, Theorem 3.6])**.** *Let (S, d) be a pointed metric space. Then, the map $\iota : S \to \mathcal{M}$ defined by $\iota(s) = m_{s0}$, for $s \in S$, isometrically embeds S into $Æ(S)$. Moreover, for every $f \in \mathrm{Lip}_0(S; Y)$ there exists a unique continuous linear map $T_f \in L(Æ(S); Y)$ such that $T_f \circ \iota = f$ with $\|T_f\|_{L(Æ(S);Y)} = \|f\|_{\mathrm{Lip}_0(S;Y)}$.*

Corollary 3.65. *Assume that (S,d) be a pointed compact metric space. Then, $(\text{Æ}(S)\hat{\otimes}_\pi Y^*)^*$ is isometrically isomorphic to $\text{Lip}_0(S;Y)$, and on bounded subsets of $\text{Lip}_0(S;Y)$, the weak-$*$-topology coincides with the topology of uniform convergence.*

Proof. By Theorem 3.64, it follows that $\text{Lip}_0(S;Y)$ is isometrically isomorphic to $\text{Æ}(S)^* \otimes Y^*$, which is in turn isometrically isomorphic to $L(\text{Æ}(S);Y)$. In this case, $\text{Æ}(S)\hat{\otimes}_\pi Y^*$ is a predual of $\text{Lip}_0(S;Y)$ identified with $L(\text{Æ}(S);Y)$, because there exists a continuous dual pairing $(\text{Æ}(S)\hat{\otimes}_\pi Y^*) \times \text{Lip}_0(S;Y) \to \mathbb{R}$ defined by

$$\langle T, f\rangle := \sum_{j=1}^{n} \langle l_j \circ f, m_j\rangle,$$

for $f \in \text{Lip}_0(S;Y)$ and $T \in \text{Æ}(S)\hat{\otimes}_\pi Y^*$ of the form $T = \sum_{j=1}^n m_j \otimes l_j$, for some $m_j \in \text{Æ}(S)$ and $l_j \in Y^*$.

Now, we show that weak-$*$ convergence on the unit ball of $\text{Lip}_0(S;Y)$ implies uniform convergence. Let $(f_n)_{n\in\mathbb{N}} \subset \text{Lip}_0(S;Y)$ with $\|f_n\|_{\text{Lip}_0(S;Y)} \leqslant 1$ converging in the weak-$*$-topology to $f \in \text{Lip}_0(S;Y)$ with $\|f\|_{\text{Lip}_0(S;Y)} \leqslant 1$. Then, it follows for $f \mapsto T(f) := \langle l \circ f, m_{s0}\rangle \in \text{Æ}(S)\hat{\otimes}_\pi Y^*$ that

$$\begin{aligned} l(f_n(s)) = l(f_n(s)) - l(f_n(0)) &= \langle T, f_n\rangle \\ &\overset{n\to\infty}{\longrightarrow} \quad \langle T, f\rangle = l(f(s)) - l(f(0)) = l(f(s)), \end{aligned} \tag{3.6.2}$$

where we have used that $f_n(0) = f(0) = 0$ implies $l(f_n(0)) = l(f(0)) = 0$. Since (3.6.2) holds true for all $l \in Y^*$ and $s \in S$, we conclude that $f_n(s) \to f(s)$, as $n \to \infty$, for all $s \in S$, showing pointwise convergence. Moreover, fix some $\varepsilon > 0$ and $n_0 \in \mathbb{N}$, and let $\{s_1, ..., s_k\}$ be an ε-net in S such that $\|f_n(s_i) - f(s_i)\|_Y < \varepsilon/3$, for all $i = 1, ..., k$ and $n \geqslant n_0$. As every $s \in S$ is at most ε away of some point in $\{s_1, ..., s_k\}$, we have

$$\begin{aligned} \|f_n(s) - f(s)\|_Y &\leqslant \|f_n(s) - f_n(s_i)\|_Y + \|f_n(s_i) - f(s_i)\|_Y + \|f(s_i) - f(s)\|_Y \\ &\leqslant d_S(s, s_i) + \|f_n(s_i) - f(s_i)\|_Y + d_S(s_i, s) < \varepsilon, \end{aligned}$$

for all $s \in S$ and some $i = 1, ..., k$, which yields uniform convergence.

Conversely, let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\text{Lip}_0(S;Y)$ with $\|f_n\|_{\text{Lip}_0(S;Y)} \leqslant 1$ converging uniformly to $f \in \text{Lip}_0(S;Y)$ with $\|f\|_{\text{Lip}_0(S;Y)} \leqslant 1$. Then, f_n converges pointwise to f, and for every $m \in \mathcal{M}$ and $l \in Y^*$, we have

$$\langle l \circ f_n, m\rangle = \sum_{s\in S} l(f_n(s))m(s) \quad \overset{n\to\infty}{\longrightarrow} \quad \sum_{s\in S} l(f(s))m(s) = \langle l \circ f, m\rangle. \tag{3.6.3}$$

Since $\text{Æ}(S)$ is defined as the $\|\cdot\|_{\text{Æ}(S)}$-closure of $\mathcal{M}$ and $\sup_{n\in\mathbb{N}} \|f_n\|_{\text{Lip}_0(S;Y)} \leqslant 1$,

(3.6.3) holds also for every $m \in Æ(S)$. In this case, (3.6.3) is still valid for $T = \sum_{k=1}^{n} m_j \otimes l_j$, with $l_j \in Y^*$ and $m_j \in \mathcal{M}$, such that $\langle T, f_n \rangle \to \langle T, f \rangle$, as $n \to \infty$, for all $T \in Æ(S)\hat{\otimes}_\pi Y^*$, which shows convergence in the weak-$*$-topology. □

Hence, we have found for a compact metric space (S, d_S) a predual of $\mathrm{Lip}_0(S;Y)$ such that $\mathrm{Lip}_0(S;Y)$ carries a weak-$*$-topology, which coincides on bounded subsets with the topology of uniform convergence.

3.6.2 Signature of Geometric Rough Paths

The predual of Lipschitz continuous functions can be used in the setting of rough paths. For this purpose, we recall some basic notation and definitions from rough path theory as used in [FV10], and introduce the free nilpotent Lie group $G^N(\mathbb{R}^d)$ in the following. The *extended tensor algebra* $T((\mathbb{R}^d))$ over $\mathbb{R}^d$ is defined as

$$T((\mathbb{R}^d)) := \bigoplus_{n=0}^{\infty} (\mathbb{R}^d)^{\otimes n},$$

where $(\mathbb{R}^d)^{\otimes n}$ denotes the nth tensor space of $\mathbb{R}^d$ with convention $(\mathbb{R}^d)^{\otimes 0} := \mathbb{R}$. On $T(\mathbb{R}^d)$, we define the standard addition "+", the tensor product "$\otimes$" used for multiplication, and scalar multiplication "$\cdot$" by

$$a + b = (a^n + b^n)_{n=0}^{\infty}, \quad a \otimes b = \left(\sum_{k=0}^{n} a^{n-k} \otimes b^k \right)_{n=0}^{\infty}, \quad \lambda \cdot a = (\lambda a^n)_{n=0}^{\infty},$$

for $a, b \in T(\mathbb{R}^d)$ and $\lambda \in \mathbb{R}$, where a^n denotes the projection to the nth level of $a \in T(\mathbb{R}^d)$. In addition, let

$$T(\mathbb{R}^d) := \left\{ a \in T((\mathbb{R}^d)); \, \exists N \in \mathbb{N} : a^n = 0, \text{ for all } n \geqslant N \right\} \subset T((\mathbb{R}^d))$$

be the *tensor algebra* over $\mathbb{R}^d$. Let $\{e_1, ..., e_d\}$ be a basis for $\mathbb{R}^d$ with dual basis $\{e_1^*, ..., e_d^*\}$ of $(\mathbb{R}^d)^*$ such that $\{e_{i_1} \otimes \cdots \otimes e_{i_n}; \, i_1, ..., i_n \in \{1, ..., d\}\}$ and $\left\{e_{i_1}^* \otimes \cdots \otimes e_{i_n}^*; \, i_1, ..., i_n \in \{1, ..., d\}\right\}$ are bases of $(\mathbb{R}^d)^{\otimes n}$ and $((\mathbb{R}^d)^*)^{\otimes n}$, respectively. Then, we observe that

$$\{e_I := e_{i_1} \otimes \cdots \otimes e_{i_n}; \, n \in \mathbb{N}, \, I = (i_1, ..., i_n) \in \{1, ..., d\}^n\} \cup \{\mathbf{1}\},$$
$$\left\{e_I^* := e_{i_1}^* \otimes \cdots \otimes e_{i_n}^*; \, n \in \mathbb{N}, \, I = (i_1, ..., i_n) \in \{1, ..., d\}^n\right\} \cup \{\mathbf{1}^*\}$$

are bases of $T(\mathbb{R}^d)$ and $T((\mathbb{R}^d)^*)$, respectively, where $\mathbf{1} = (1, 0, 0, ...) \in T(\mathbb{R}^d)$ is the unit element, and where $\mathbf{1}^* \in T((\mathbb{R}^d)^*)$ is such that $\mathbf{1}^*(a) = a^0$, for all $a \in T(\mathbb{R}^d)$.

Definition 3.66. *For $I = (i_1, ..., i_n) \in \{1, ..., d\}^n$ and $J = (j_1, ..., j_m) \in \{1, ..., d\}^m$, we define the* shuffle product $\sqcup\!\sqcup$ *inductively by*

$$e_I^* \sqcup\!\sqcup e_J^* := \left(e_{(i_1,...,i_{n-1})}^* \sqcup\!\sqcup e_J^*\right) \otimes e_{i_n}^* + \left(e_I^* \sqcup\!\sqcup e_{(j_1,...,j_{m-1})}^*\right) \otimes e_{j_m}^*,$$

with $e_I^* \sqcup\!\sqcup \mathbf{1}^* = \mathbf{1}^* \sqcup\!\sqcup e_I^* = e_I^*$, which is extended by bilinearity to the extended tensor algebra $T((\mathbb{R}^d)^*)$.

Let $C^{1-\mathrm{var}}([0,T];\mathbb{R}^d)$ be the space of continuous paths $Z : [0,T] \to \mathbb{R}^d$ of finite variation. For a given $N \in \mathbb{N}$ and a path $Z \in C^{1-\mathrm{var}}([0,T];\mathbb{R}^d)$, we define its *step-$N$ signature* over the interval $[s,t] \subset [0,T]$ by

$$S_N(Z)_{s,t} := \left(1, \int_{s<u<t} dZ_u, ..., \int_s^t \int_s^{u_N} \cdots \int_s^{u_2} dZ_{u_1} \otimes \cdots \otimes dZ_{u_{N-1}} \otimes dZ_{u_N}\right) \in T^N(\mathbb{R}^d),$$

where $T^N(\mathbb{R}^d) := \bigoplus_{n=0}^N (\mathbb{R}^d)^{\otimes n}$ denotes the truncated tensor algebra of order N over $\mathbb{R}^d$. The corresponding space of lifted paths of finite variation in $C^{1-\mathrm{var}}([0,T];\mathbb{R}^d)$ is the step-N free nilpotent group

$$G^N(\mathbb{R}^d) := \left\{S_N(Z)_{0,T};\ Z \in C^{1-\mathrm{var}}([0,T];\mathbb{R}^d)\right\} \subset T^N(\mathbb{R}^d),$$

which is equipped with one of the following metrics, either the submanifold metric $d^1_{G^N(\mathbb{R}^d)}(g,h) = \max_{k=1,...,N} \left\|g^k - h^k\right\|$ or the Carnot-Carathéodory metric (see [FV10, Definition 7.41]) defined by $d^2_{G^N(\mathbb{R}^d)}(g,h) = \left\|g^{-1} \otimes h\right\|_{cc}$, where $\|g\|_{cc} := \inf\left\{\int_0^T |dZ|;\ Z \in C^{1-\mathrm{var}}([0,T];\mathbb{R}^d),\ S_N(Z)_{0,T} = g\right\}$. These metrics are not equivalent (unless $d = 1$), but there exists by [FV10, Proposition 7.49] some $C > 0$ such that

$$d^1_{G^N(\mathbb{R}^d)}(g,h) \leqslant C d^2_{G^N(\mathbb{R}^d)}(g,h), \qquad d^2_{G^N(\mathbb{R}^d)}(g,h) \leqslant C d^1_{G^N(\mathbb{R}^d)}(g,h)^{1/N},$$

for all $g, h \in G^N(\mathbb{R}^d)$ belonging to a bounded subset of $G^N(\mathbb{R}^d)$ with respect to $\|\cdot\|_{cc}$.

In the following, we use the Carnot-Carathéodory metric and introduce for a path $\mathbb{X} : [0,T] \to G^N(\mathbb{R}^d)$ the map $\mathbb{X} : \Delta_2 \to G^N(\mathbb{R})$ defined by $\mathbb{X}_{s,t} = \mathbb{X}_s^{-1} \otimes \mathbb{X}_t$, for $(s,t) \in \Delta_2$, where $\Delta_2 = \left\{(s,t) \in [0,T]^2;\ s \leqslant t\right\}$ denotes the two-dimensional simplex. For $0 < \alpha \leqslant 1$, we introduce the α-Hölder metric over

$[s,t]$ given by

$$d_{\alpha-\text{Höl},[s,t]}(\mathbb{X},\mathbb{Y}) = \max_{n=1,\dots,\lfloor 1/\alpha\rfloor}\left(\sup_{\substack{u,v\in[s,t]\\u<v}}\frac{\|\mathbb{X}^n_{u,v}-\mathbb{Y}^n_{u,v}\|}{|u-v|^{n\alpha}}\right)^{\frac{1}{n}},$$

for $\mathbb{X},\mathbb{Y}:\Delta_2\to G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)$. Using this, we define the space of geometric α-Hölder rough paths as the $d_{\alpha-\text{Höl},[0,T]}$-closure of the step-$N$ signatures obtained from paths with finite variation.

Definition 3.67. *For $0<\alpha\leqslant 1$, a map $\mathbb{X}:\Delta_2\to G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)$ is a* geometric α-Hölder rough path *if there exists a sequence $(X(k))_{k\in\mathbb{N}}\subset C^{1-var}([0,T];\mathbb{R}^d)$ such that $(S_{\lfloor 1/\alpha\rfloor}(X(k)))_{k\in\mathbb{N}}:\Delta_2\to G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d)$ converges in the $d_{\alpha-Höl,[0,T]}$-metric to $\mathbb{X}$. We denote by $G\Omega_\alpha([0,T];\mathbb{R}^d)$ the space of geometric α-Hölder rough paths.*

In addition, every geometric α-Hölder rough path $\mathbb{X}\in G\Omega_\alpha([0,T];\mathbb{R}^d)$ can be uniquely extended to a map $\mathbb{X}^{<\infty}:\Delta_2\to T(\mathbb{R}^d)$ by [LCL07, Theorem 3.7], which we call its *signature*. For example, the sample paths of a continuous semimartingale are almost surely a geometric α-Hölder rough path with $\alpha\in\,]1/3,1/2]$. In addition, the shuffle product can be used to describe the product of two linear functionals.

Lemma 3.68 ([LCL07, Theorem 2.15])**.** *Let $\mathbb{X}\in G\Omega_\alpha([0,T];\mathbb{R}^d)$ and $l_1,l_2\in T((\mathbb{R}^d)^*)$. Then, it holds that*

$$l_1\left(\mathbb{X}^{<\infty}_{s,t}\right)\,l_2\left(\mathbb{X}^{<\infty}_{s,t}\right)=(l_1\sqcup\!\sqcup l_2)\left(\mathbb{X}^{<\infty}_{s,t}\right),$$

for all $(s,t)\in\Delta_2$.

By Lemma 3.68, every $\mathbb{X}\in G\Omega_\alpha([0,T];\mathbb{R}^d)$ satisfies Chen's relation $\mathbb{X}^{<\infty}_{s,u}\otimes\mathbb{X}^{<\infty}_{u,t}=\mathbb{X}^{<\infty}_{s,t}$, for all $0\leqslant s\leqslant u\leqslant t\leqslant T$, which means that the whole path information is contained in its signature $\mathbb{X}^{<\infty}_{0,T}$. In the following, we consider continuous paths $X:[0,T]\to\mathbb{R}^d$, for example a sample path of a continuous semimartingale. Hereby, we assume $X(0)=0$ and define the space of rough paths $\Omega_T=\left\{t\mapsto\mathbb{X}_t:=\mathbb{X}_{0,t}:[0,T]\to G^{\lfloor 1/\alpha\rfloor}(\mathbb{R}^d);\,\mathbb{X}\in G\Omega_\alpha([0,T];\mathbb{R}^d),\,X(0)=0\right\}$, which we equip with the seminorm

$$\|\mathbb{X}\|_\alpha=\max_{n=1,\dots,\lfloor 1/\alpha\rfloor}\left(\sup_{\substack{s,t\in[0,T]\\s<t}}\frac{\|\mathbb{X}^n_t-\mathbb{X}^n_s\|}{|t-s|^{n\alpha}}\right)^{\frac{1}{n}},$$

where $\|\cdot\|$ denotes the norm on $(\mathbb{R}^d)^{\otimes n}$.

Definition 3.69. *For $X \in C^0([0,T];\mathbb{R}^d)$, we define its* time-augmentation *as $t \mapsto \widehat{X}_t = (t, X_t) \in \mathbb{R}^{d+1}$. From this, we define the rough path space $\widehat{\Omega}_T = \left\{ t \mapsto \widehat{\mathbb{X}}_t := \widehat{\mathbb{X}}_{0,t} : [0,T] \to T(\mathbb{R}^{d+1});\ \mathbb{X} \in G\Omega_\alpha([0,T];\mathbb{R}^d),\ X(0) = 0 \right\}$.*

We now show that every continuous functional depending on $\mathbb{X} \in \widehat{\Omega}_T$ can be approximated with the help of linear functionals in $T((\mathbb{R}^{d+1})^*)$ applied on the signature $\mathbb{X}_{0,T}^{<\infty}$. For this purpose, we introduce the weight function $\psi : \widehat{\Omega}_T \to]0,\infty[$ given by

$$\psi(\widehat{\mathbb{X}}) = \exp\left(\|\widehat{\mathbb{X}}\|_\infty^2 \right), \tag{3.6.4}$$

for $\widehat{\mathbb{X}} \in \widehat{\Omega}_T$, where $\|\widehat{\mathbb{X}}\|_\infty = \sup_{t\in[0,T]} \max_{n=1,\dots,\lfloor 1/\alpha \rfloor} \|\widehat{\mathbb{X}}_t^n\|$. Using the results of Section 3.6.1, we know that $\widehat{\Omega}_T$ carries a weak-$*$-topology, where every functional in $\mathcal{B}_\psi(\widehat{\Omega}_T)$ can be approximated by a linear combination of the signature $\widehat{\mathbb{X}}_{0,T}^{<\infty} \in T(\mathbb{R}^{d+1})$, for which we first show the following lemma.

Lemma 3.70. *Let $\mathbb{X} \in G\Omega_\alpha([0,T];\mathbb{R}^d)$ and assume that there exists some $l \in T^{\lfloor 1/\alpha \rfloor}((\mathbb{R}^d)^*)$ such that $t \mapsto l(\mathbb{X}_{0,t})$ is strictly monotone. Then, the signature $\mathbb{X}_{0,T}^{<\infty}$ uniquely determines $\mathbb{X} : \Delta_2 \to G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d)$.*

Proof. This follows immediately from [BGLY16], where the signature $\mathbb{X}_{0,T}^{<\infty}$ uniquely determines $\mathbb{X} : [0,T] \to G^{\lfloor 1/\alpha \rfloor}(\mathbb{R}^d)$ up to tree-like equivalences (see [BGLY16, Definition 1.1]). However, they are ruled out by assuming that there exists a linear functional $l \in T^{\lfloor 1/\alpha \rfloor}((\mathbb{R}^d)^*)$ such that $t \mapsto l(\mathbb{X}_{0,t})$ is strictly monotone. □

Theorem 3.71. *The set $\left\{ \widehat{\Omega}_T \ni \widehat{\mathbb{X}} \mapsto l(\widehat{\mathbb{X}}_{0,T}^{<\infty});\ l \in T((\mathbb{R}^{d+1})^*) \right\}$ is τ_ψ^0-dense in $\mathcal{B}_\psi(\widehat{\Omega}_T)$, i.e. for every $f \in \mathcal{B}_\psi(\widehat{\Omega}_T)$ and $\varepsilon > 0$ there exists some $l \in T((\mathbb{R}^{d+1})^*)$ such that*

$$\sup_{\widehat{\mathbb{X}} \in \widehat{\Omega}_T} \frac{\left| f(\widehat{\mathbb{X}}) - l(\widehat{\mathbb{X}}_{0,T}^{<\infty}) \right|}{\psi(\widehat{\mathbb{X}})} < \varepsilon.$$

Proof. For the Banach space $Y = \bigoplus_{n=1}^N (\mathbb{R}^d)^{\otimes n}$ with norm $\max_{n=1,\dots,N} \|a^k\|$, we observe that $\widehat{\Omega}_T$ contains Y-valued Lipschitz functions $t \mapsto \mathbb{X}_{0,t} : [0,T] \to Y$ with $\mathbb{X}(0) = 0$ defined on the compact metric space $([0,T], d_\alpha)$ with $d_\alpha(s,t) =$

$|t-s|^\alpha$. Since $(Y, \|\cdot\|_Y)$ is finite dimensional, the space $Æ(S)\hat{\otimes}_\pi Y^*$ is by Corollary 3.65 a predual of $\mathrm{Lip}_0(S;Y)$, where the weak-$*$-topology coincides with the topology of uniform convergence on bounded subsets of $\widehat{\Omega}_T$. From this, we define the weight function $\psi : \widehat{\Omega}_T \to]0,\infty[$ as in (3.6.4) such that the preimage $K_R = \psi^{-1}(]0,R])$ is bounded and by Banach-Alaoglu compact in the weak-$*$-topology, for all $R > 0$, and $\psi : \widehat{\Omega}_T \to]0,\infty[$ therefore an admissible weight function. Moreover, we observe that $\widehat{\mathbb{X}} \mapsto l_1(\widehat{\mathbb{X}}^{<\infty}_{0,T}) + l_2(\widehat{\mathbb{X}}^{<\infty}_{0,T}) = (l_1 + l_2)(\widehat{\mathbb{X}}^{<\infty}_{0,T}) \in \mathcal{A}$ and $\widehat{\mathbb{X}} \mapsto l_1(\widehat{\mathbb{X}}^{<\infty}_{0,T})\, l_2(\widehat{\mathbb{X}}^{<\infty}_{0,T}) = (l_1 ⧢ l_2)(\widehat{\mathbb{X}}^{<\infty}_{0,T}) \in \mathcal{A}$, by Lemma 3.68, for all $l \in T((\mathbb{R}^{d+1})^*)$ and $\widehat{\mathbb{X}} \in \widehat{\Omega}_T$, such that

$$\mathcal{A} = \left\{ \widehat{\Omega}_T \ni \widehat{\mathbb{X}} \to l(\widehat{\mathbb{X}}^{<\infty}_{0,T});\, l \in T((\mathbb{R}^{d+1})^*) \right\}$$

is a subalgebra of $\mathcal{B}_\psi(\widehat{\Omega}_T)$, see also [LCL07, Theorem 2.15]. In addition, it follows with Lemma 3.70 and the time augmentation that the signatures $\mathbb{X}^{<\infty}_{0,T}$ and $\mathbb{Y}^{<\infty}_{0,T}$ of distinct $\mathbb{X}, \mathbb{Y} \in \widehat{\Omega}_T$ differ, which implies for some $l \in T((\mathbb{R}^{d+1})^*)$ that the function $\widehat{\mathbb{X}} \mapsto a(\widehat{\mathbb{X}}^{<\infty}) := l(\widehat{\mathbb{X}}^{<\infty}_{0,T}) \in \mathcal{A}$ satisfies $a(\widehat{\mathbb{X}}) := l(\widehat{\mathbb{X}}^{<\infty}_{0,T}) \neq l(\widehat{\mathbb{Y}}^{<\infty}_{0,T}) = a(\widehat{\mathbb{Y}})$. Hence, $\mathcal{A}$ is locally point separating with point separating subspace $\widetilde{\mathcal{A}} = \mathcal{A}$ because

$$\begin{aligned}\sup_{\widehat{\mathbb{X}}\in\widehat{\Omega}_T} \frac{\exp\left(\left|a(\widehat{\mathbb{X}})\right|\right)}{\psi(\widehat{\mathbb{X}})} &= \sup_{\widehat{\mathbb{X}}\in\widehat{\Omega}_T} \frac{\exp\left(\left|l(\widehat{\mathbb{X}}^{<\infty}_{0,T})\right|\right)}{\psi(\widehat{\mathbb{X}})} \\ &\leqslant \sup_{\widehat{\mathbb{X}}\in\widehat{\Omega}_T} \frac{\exp\left(\sum_{n=1}^N \|l_n\|_{((\mathbb{R}^{d+1})^*)^{\otimes n}} \|\widehat{\mathbb{X}}^n_{0,T}\|\right)}{\exp\left(\sup_{t\in[0,T]} \max_{n=1,\dots,\lfloor 1/\alpha\rfloor} \|\widehat{\mathbb{X}}^n_{0,t}\|^2\right)} < \infty,\end{aligned}$$

for all $a \in \mathcal{A}$ of the form $a(\widehat{\mathbb{X}}) = l(\widehat{\mathbb{X}}^{<\infty}_{0,T}) = \sum_{n=0}^N l_n(\widehat{\mathbb{X}}^n_{0,T})$, where $l \in T((\mathbb{R}^{d+1})^*)$ can be written as $l = \sum_{n=0}^N l_n$, for some $l_n \in ((\mathbb{R}^{d+1})^{\otimes n})^*$. In addition, $\mathcal{A}$ vanishes nowhere because for every $\widehat{\mathbb{X}} \in \widehat{\Omega}_T$ we can choose $\mathbf{1}^* \in T((\mathbb{R}^{d+1})^*)$ such that $\widehat{\mathbb{X}} \mapsto a(\widehat{\mathbb{X}}) := \mathbf{1}^*(\widehat{\mathbb{X}}^{<\infty}_{0,T}) \in \mathcal{A}$ satisfies $a(\widehat{\mathbb{X}}) = 1$, for all $\widehat{\mathbb{X}} \in \widehat{\Omega}_T$. Then, the conclusion follows from the weighted Stone-Weierstrass in Theorem 3.7. □

In Theorem 3.71, we have used the weak-$*$-topology of $\widehat{\Omega}_T$ as weighted topology to consider the approximation beyond compact subsets of $\widehat{\Omega}_T$, which differs from the usual rough path norms. Moreover, one would only consider the first N iterated integrals for an application in practice and learn a given functional $f \in \mathcal{B}_\psi(\widehat{\Omega}_T)$ by regressing on the truncated signature $\mathbb{X}^{\leqslant N}_{0,T} = (1, \mathbb{X}^1_{0,T}, ..., \mathbb{X}^N_{0,T}) \in T^N(\mathbb{R}^d)$.

Chapter 4

Universal Approximation on Space of Stopped Paths

The space of stopped paths (Λ_T, d_∞) defined in the introductory example in Section 3.1 was originally introduced in [Dup09], [CF10] and [CF13] to generalize Föllmer's pathwise Ito calculus (see Section 1.2.4). More precisely, the pathwise Ito formula in Theorem 1.41, also called the change of variable formula, was generalized by [CF10] to non-anticipative functionals by using their horizontal and vertical derivatives. Hereby, the non-anticipative functional is able to depend on the history of the trajectory, and we introduce a special type of abstract neural networks, so-called "functional neural networks", for which we prove a particular universal approximation result with derivatives such that not only the value of the non-anticipative functional can be approximated, but also their horizontal and vertical derivatives.

4.1 Non-Anticipative Functional Calculus

In the following, we introduce non-anticipative functionals and the pathwise functional Ito calculus, where we now use capital letters for paths and functionals in contrast to Chapter 3. Let $U \subset \mathbb{R}^d$ be an open subset and consider a continuous path $Y \in C^0([0,T];U)$. We define the *stopped path* $Y^t : [0,T] \to U$ at time $t \in [0,T]$ by $Y^t(s) = Y(s \wedge t)$, for all $s \in [0,T]$, and introduce the space of stopped paths Λ_T given by

$$\Lambda_T = \{(t, Y^t);\, (t,Y) \in [0,T] \times C^0([0,T],U)\} \cong \big([0,T] \times C^0([0,T],U)\big) / \sim,$$

where $\sim$ denotes the equivalence relation defined by $(t,Y) \sim (s,Z)$ if and only if $t = s$ and $Y^t = Z^s$. Moreover, we equip Λ_T with the metric $d_\infty : \Lambda_T \times \Lambda_T \to$

$[0, \infty[$ defined by

$$d_\infty((t, Y), (s, Z)) = |t - s| + \sup_{u \in [0,T]} \|Y^t(u) - Z^s(u)\|,$$

for $(t, Y), (s, Z) \in \Lambda_T$ such that (Λ_T, d_∞) becomes a complete metric space. With this, we can now introduce path-dependent functionals, which are called non-anticipative functionals.

Definition 4.1. *A* non-anticipative functional *is a measureable map $F : \Lambda_T \to \mathbb{R}$. In addition, F is* continuous *if F is continuous with respect to d_∞, i.e. if for every $\varepsilon > 0$ there exists some $\delta > 0$ such that*

$$d_\infty((t, Y), (s, Z)) < \delta \quad \Longrightarrow \quad |F(t, Y) - F(s, Z)| < \varepsilon.$$

By $\mathbb{C}^{0,0}(\Lambda_T)$, we denote the vector space of all continuous non-anticipative functionals defined on Λ_T.

Moreover, a non-anticitpative functional $F : \Lambda_T \to \mathbb{R}$ is called *boundedness-preserving* if for every compact subset $K \subset \mathbb{R}^d$ there exists a constant $C > 0$ such that $(t, Y) \in [0, T] \times C^0([0, T]; K)/\sim$ implies that $|F(t, Y)| \leqslant C$. In addition, a non-anticipative functional $F : \Lambda_T \to \mathbb{R}$ is called *continuous at fixed times* if the map $Y \mapsto F(t, Y)$ is continuous with respect to the supremum norm $\|\cdot\|_{C^0}$, for all $t \in [0, T]$.

Next, we recall the notions of differentiability for non-anticipative functionals. A non-anticipative functional $F : \Lambda_T \to \mathbb{R}$ is called *horizontally differentiable* if the limit

$$\mathcal{D}F(t, Y^t) = \lim_{h \to 0^+} \frac{F(t + h, Y^t) - F(t, Y^t)}{h}$$

exists for all $(t, Y) \in \Lambda_T$ and if $\mathcal{D}F : \Lambda_T \to \mathbb{R}$ is a boundedness-preserving non-anticipative functional such that for every $0 \leqslant s < t \leqslant T$ and $Y \in C^0([0, T], U)$, the function $[s, t] \ni r \mapsto \mathcal{D}F(r, Y^s)$ is Borel-measurable and it holds that

$$F(t, Y^s) - F(s, Y^s) = \int_s^t \mathcal{D}F(r, Y^s) dr.$$

On the other hand, $F : \Lambda_T \to \mathbb{R}$ is called *vertically differentiable* if the map $v \mapsto F(t, Y + v\mathbb{1}_{[t,T]})$ is differentiable at 0, for all $(t, Y) \in \Lambda_T$. In this case, the vertical derivatives of F at (t, Y) are denoted by $\nabla_X F(t, Y) = (\partial_i F(t, Y))_{i=1,\dots,d}$,

for which the ith partial derivative is defined as

$$\partial_i F(t,Y) = \lim_{h\to 0} \frac{F(t, Y^t + he_i \mathbb{1}_{[t,T]}) - F(t,Y^t)}{h},$$

for $i = 1, ..., d$ and $(t,Y) \in \Lambda_T$. We denote by $\mathbb{C}^{p,q}(\Lambda_T)$ the space of non-anticipative functionals $F \in \mathbb{C}^{0,0}(\Lambda_T)$ such that $F : \Lambda_T \to \mathbb{R}$ is continuous at fixed times, p-times horizontally differentiable and q-times vertically differentiable, where $\mathcal{D}F, ..., \mathcal{D}^p F \in \mathbb{C}^{0,0}(\Lambda_T)$ and $\nabla_X F, ..., \nabla_X^q F \in \mathbb{C}^{0,0}(\Lambda_T)$ are continuous at fixed times and boundedness-preserving. Under Assumption 1.40, we obtain the following pathwise functional Ito formula, which generalizes Föllmer's pathwise Ito formula in Theorem 1.41.

Lemma 4.2 ([CF10, Theorem 3])**.** *Suppose that the path $X : [0,T] \to U$ satisfies Assumption 1.40, where $U \subset \mathbb{R}^d$ is open, and assume that $F \in \mathbb{C}^{1,2}(\Lambda_T)$. Then,*

$$F(t,X) - F(0,X) = \int_0^t \mathcal{D}F(s,X)ds + \int_0^t \nabla_X F(s,X)^\top dX(s) + \frac{1}{2}\sum_{i,j=1}^d \int_0^t \partial_{ij} F(s,X) d[X_i, X_j]_s,$$

for all $t \in [0,T]$, where the pathwise stochastic integral is defined as

$$\int_0^t \nabla_X F(s,X)^\top dX(s) = \lim_{n\to\infty} \sum_{s\in\mathbb{T}_n, s\leqslant t} \nabla_X F(s,X)^\top \left(X(s') - X(s)\right), \quad (4.1.1)$$

for $t \in [0,T]$, and where the series on the right-hand side of (4.1.1) *converges absolutely.*

In addition, a functional $\xi : [0,T] \times C([0,T],U) \to \mathbb{R}^d$ is called an *admissible functional integrand* for Y on U if there exists some $F \in \mathbb{C}_b^{1,2}(\Lambda_T)$ such that $\xi(t,Y) = \nabla_X F(t,Y)$, for all $(t,Y) \in [0,T] \times C([0,T],U)$, where $U \subset \mathbb{R}^d$ is open. In this case, Lemma 4.2 gives a pathwise expression for the stochastic integral, which we call the *Föllmer integral* as in Section 1.2.4.

4.2 Functional Neural Networks

Similarly as in Chapter 3, we would like to approximate a non-anticipative functional by a specific type of neural network such that the non-anticipative behaviour is preserved. Following the ideas of Section 3.3, we construct an

additive family between Λ_T and $\mathbb{R}$ in order to define functional neural network as follows.

Definition 4.3. *Let $N, M_1 \in \mathbb{N}_{\geqslant 1}$ and $\rho : \mathbb{R} \to \mathbb{R}$ be a function such that $\varphi_{ki} \in \mathcal{NN}^{\rho}_{d,M_1}$ are conventional neural networks of connectivity M_1, for $k = 1, ...N$ and $i = 1, ..., d$. Then, the map $\Phi : \Lambda_T \to \mathbb{R}$ defined by*

$$\Phi(t, Y) = \sum_{k=1}^{N} \beta_k \rho \left(\sum_{i=1}^{d} \int_0^t \varphi_{ki}(u) Y_i(u) du + \sum_{i=1}^{d} A_{ki} Y_i(t) + B_k t + b_k \right), \quad (4.2.1)$$

for $(t, Y) \in \Lambda_T$, is called a functional neural network (FNN)*, where $N \in \mathbb{N}_{\geqslant 1}$ is the number of neurons, $A \in \mathbb{R}^{N \times d}$ and $B \in \mathbb{R}^N$ are* weight matrices*, $b \in \mathbb{R}^N$ is the* bias vector*, and where $\beta \in \mathbb{R}^N$ denotes the* linear readout*. Moreover, we denote by $M \in \mathbb{N}$ the number of non-zero entries in A and B, for which the connectivity of the neural networks $(\varphi_{ki})_{k=1,...,N, i=1,...,d}$ is added, and call it the network's* connectivity*.*

By construction, a functional neural network $\Phi : \Lambda_T \to \mathbb{R}$ satisfies $\Phi(t, Y) = \Phi(t, Y^t)$, for all $(t, Y) \in \Lambda_T$, such that $\Phi : \Lambda_T \to \mathbb{R}$ is a non-anticipative functional. Moreover, if $\rho \in C^0(\mathbb{R})$ then $\Phi \in \mathbb{C}^{0,0}(\Lambda_T)$, whereas $\rho \in C^2_b(\mathbb{R})$ implies $\Phi \in \mathbb{C}^{1,2}(\Lambda_T)$. In the latter case, the horizontal and vertical derivatives are given by

$$\begin{aligned}
\mathcal{D}\Phi(t, Y) &= \sum_{k=1}^{N} \beta_k \rho' \left(\sum_{i=1}^{d} \int_0^t \varphi_{ki}(u) Y_i(u) du + A_{k\cdot}^{\top} Y(t) + B_k t + b_k \right) \cdots \\
&\qquad \cdots \left(\sum_{i=1}^{d} \varphi_{ki}(t) Y_i(t) + B_k \right), \\
\partial_l \Phi(t, Y) &= \sum_{k=1}^{N} \beta_k \rho' \left(\sum_{i=1}^{d} \int_0^t \varphi_{ki}(u) Y_i(u) du + A_{k\cdot}^{\top} Y(t) + B_k t + b_k \right) A_{kl}, \\
\partial_{lm} \Phi(t, Y) &= \sum_{k=1}^{N} \beta_k \rho'' \left(\sum_{i=1}^{d} \int_0^t \varphi_{ki}(u) Y_i(u) du + A_{k\cdot}^{\top} Y(t) + B_k t + b_k \right) A_{kl} A_{km},
\end{aligned}$$

for all $l, m = 1, ..., d$ and $(t, Y) \in \Lambda_T$. By $\mathcal{FNN}^{\rho}_{d,M}$, we denote the set of functional neural networks of at most connectivity M, and let $\mathcal{FNN}^{\rho}_{d,\infty} = \bigcup_{M \in \mathbb{N}} \mathcal{FNN}^{\rho}_{d,M}$ be the set of all FNNs with arbitrary connectivity.

Remark 4.4. *Deep FNNs can be constructed by concatenation. Because* $\mathcal{FN}^{\rho}_{d,\infty}$ *is again an additive family between* Λ_T *and* $\mathbb{R}$ *in the sense of Definition 3.13, the non-anticipative functional* $\Phi : \Lambda_T \to \mathbb{R}$ *given by*

$$\Phi(t,Y) = \sum_{k=1}^{N} \beta_k \rho\left(\Psi_k(t,Y) + b_k\right),$$

for $(t,Y) \in \Lambda_T$*, is a functional neural networks with two hidden layers, where* $\beta, b \in \mathbb{R}^N$ *and* $\Psi_k \in \mathcal{FN}^{\rho}_{d,\infty}$*.*

4.3 Universal Approximation with Derivatives

In the following, we use the universal approximation result in Theorem 3.52 to show that every smooth functional including its derivatives can be approximated arbitrarily well by a functional neural network. For this purpose, we restrict ourselves to the space of stopped α-Hölder continuous paths Λ^{α}_T introduced in Example 3.2(v), which is defined as elements $(t,Y) \in \Lambda_T$ with path Y having finite Hölder norm

$$\|Y\|_{C^{\alpha}([0,T];\mathbb{R}^d)} = \sup_{s\in[0,T]} \|Y(s)\| + \sup_{u\neq s} \frac{\|Y(u) - Y(s)\|}{|u-s|^{\alpha}}.$$

Using this, we define the weight function $\psi : \Lambda^{\alpha}_T \to]0,\infty[$ by $\psi(t,Y) = \exp\left(\|Y\|_{C^{\alpha}([0,T];\mathbb{R}^d)}\right)$, for $(t,Y) \in \Lambda^{\alpha}_T$ and some fixed $\alpha > 0$, such that the preimage $K_R = \psi^{-1}(]0,R])$ is equal to $K_R = ([0,T] \times A_R)/\sim$, for some equicontinuous and bounded subset $A_R \subset C^0([0,T];\mathbb{R}^d)$. Hence, A_R is by Arzelà-Ascoli in Theorem 1.17 compact with respect to $\|\cdot\|_{C^0([0,T];\mathbb{R}^d)}$ and $[0,T] \times A_R$ is by Tychonoff therefore compact in the product topology of $[0,T] \times C^0([0,T];\mathbb{R}^d)$ such that K_R is compact in the quotient topology induced by d_{∞}. Because this holds for every $R > 0$, the weight function $\psi : \Lambda^{\alpha}_T \to]0,\infty[$ is admissible in the sense of Definition 3.1.

Remark 4.5. *Using the weight function $\psi : \Lambda_T^\alpha \to]0,\infty[$ with $\alpha < 1/2$, we conclude for a path $Y : [0,T] \to \mathbb{R}^d$ satisfying Assumption 1.40 that (t, Y^t) belongs to Λ_T^α, because there exists some $R > 0$ such that*

$$\begin{aligned}\psi(t,Y) &= \exp\left(\sup_{s\in[0,T]}\|Y(s)\| + \sup_{u\neq s}\frac{\|Y(u)-Y(s)\|}{|u-s|^\alpha}\right)\\ &\leqslant \exp\left(\|Y\|_{C^0([0,T];\mathbb{R}^d)} + \max_{i=1,\dots,d}[Y_i,Y_i]_T^{\frac{1}{2}-\alpha}\right) \leqslant R,\end{aligned}$$

for all $t \in [0,T]$. This means in particular that almost every stopped path of a sample path of a d-dimensional Brownian motion $B = (B(t))_{t\in[0,T]}$ defined on a filtered probability space $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{P})$ belongs to Λ_T^α.

We now restrict ourselves to the space of smooth functionals $\mathbb{C}^{p,q}(\Lambda_T^\alpha)$ defined on the space of stopped α-Hölder continuous paths Λ_T^α. In addition, we introduce the vector space $\mathcal{L}\mathrm{ip}_b^{p,q}(\Lambda_T^\alpha)$ of bounded and smooth $\mathcal{L}\mathrm{ip}^{p,q}$-functionals with bounded horizontal derivative up to order p and bounded vertical derivatives up to order q, where the highest derivatives are locally Lipschitz continuous. We equip $\mathcal{L}\mathrm{ip}_b^{p,q}(\Lambda_T^\alpha)$ with the norm

$$\begin{aligned}\|F\|_{p,q} = &\sup_{(t,Y)\in\Lambda_T^\alpha}|F(t,Y)| + \sum_{j=1}^{p}\sup_{(t,Y)\in\Lambda_T^\alpha}\left|\mathcal{D}^jF(t,Y)\right|\\ &+ \sum_{1\leqslant|\mathbf{m}|\leqslant q}\sup_{(t,Y)\in\Lambda_T^\alpha}|\partial_{\mathbf{m}}F(t,Y)|.\end{aligned}$$

Furthermore, we introduce the set $B_\psi^{p,q}(\Lambda_T^\alpha) = \{F : \Lambda_T^\alpha \to \mathbb{R};\ \|F\|_{\psi,p,q} < \infty\}$ of non-anticipative functionals, whose growth as well as the growth of their horizontal and vertical derivatives are controlled by the growth of $\psi : \Lambda_T^\alpha \to]0,\infty[$, where $\|\cdot\|_{\psi,p,q}$ denotes the weighted norm defined by

$$\begin{aligned}\|F\|_{\psi,p,q} = &\sup_{(t,Y)\in\Lambda_T^\alpha}\frac{|F(t,Y)|}{\psi(t,Y)} + \sum_{j=1}^{p}\sup_{(t,Y)\in\Lambda_T^\alpha}\frac{\left|\mathcal{D}^jF(t,Y)\right|}{\psi(t,Y)}\\ &+ \sum_{1\leqslant|\mathbf{m}|\leqslant q}\sup_{(t,Y)\in\Lambda_T^\alpha}\frac{|\partial_{\mathbf{m}}F(t,Y)|}{\psi(t,Y)}.\end{aligned}$$

Finally, we define $\mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$ as the completion of $\mathcal{L}\mathrm{ip}_b^{p,q}(\Lambda_T^\alpha)$ with respect to the weighted norm $\|\cdot\|_{\psi,p,q}$, where it follows for some $F \in \mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$ similarly

as in the proof of Lemma 3.4 with $K_R = \psi^{-1}(]0,R])$ that

$$\lim_{R\to\infty}\left(\sup_{(t,Y)\in\Lambda_T^\alpha\setminus K_R}\frac{|F(t,Y)|}{\psi(t,Y)}+\sum_{j=1}^{p}\sup_{(t,Y)\in\Lambda_T^\alpha\setminus K_R}\frac{\left|\mathcal{D}^jF(t,Y)\right|}{\psi(t,Y)}+\cdots\right.$$
$$\left.\cdots+\sum_{1\leqslant|\mathbf{m}|\leqslant q}\sup_{(t,Y)\in\Lambda_T^\alpha\setminus K_R}\frac{|\partial_{\mathbf{m}}F(t,Y)|}{\psi(t,Y)}\right)=0,$$

where we can use that every $F\in\mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$ can be approximated by cylindrical functionals such that the topology induced by $\|\cdot\|_{\psi,p,q}$ is weaker than the weighted topology $\tau_\psi^{\max(p,q)}$ for the cylindrical functionals.

We now apply the universal approximation result with derivatives derived in Chapter 3 to show that every smooth functionals $\mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$ can be approximated arbitrarily well by a functional neural network. As activation function, we consider a $\mathcal{L}\mathrm{ip}^{\max(p,q)}$-function $\rho\in\mathcal{L}\mathrm{ip}_b^{\max(p,q)}(\mathbb{R})$ such that

$$\mathcal{NN}_{\mathbb{R},\mathbb{R}}^{\mathcal{G},\{\rho\},\mathcal{K}}:=\left\{\mathbb{R}\ni x\mapsto\sum_{k=1}^{N}\beta_k\rho\left(a_kx+b_k\right)\in\mathbb{R};\ N\in\mathbb{N}_{\geqslant1},\ a_k,b_k,\beta_k\in\mathbb{R}\right\}\tag{4.3.1}$$

is dense in $C^{\max(p,q)}(K)$, for all compact subsets $K\subset\mathbb{R}$, where we have $\mathcal{G}=\{z\mapsto az+b;\ a,b\in\mathbb{R}\}$ and $\mathcal{K}=\{z\mapsto wz;\ w\in\mathbb{R}\}$. For this purpose, Remark 3.50 gives sufficient conditions on $\rho\in\mathcal{L}\mathrm{ip}_b^{\max(p,q)}(\mathbb{R})$ such that $\mathcal{NN}_{\mathbb{R},\mathbb{R}}^{\mathcal{G},\{\rho\},\mathcal{K}}$ is $\tau_{\psi_{\mathbb{R}}}^{\max(p,q)}$-dense in $\mathcal{L}\mathrm{ip}_{\psi_{\mathbb{R}}}^{\max(p,q)}(\mathbb{R})$, where $\psi_{\mathbb{R}}:\mathbb{R}\to]0,\infty[$ is an admissible weight function on $\mathbb{R}$.

Theorem 4.6 (Universal Approximation on $\mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$)**.** *Assume that $\rho\in\mathcal{L}ip_b^{\max(p,q)}(\mathbb{R})$ is such that (4.3.1) is dense in $C^{\max(p,q)}(K)$, for all compact subsets $K\subset\mathbb{R}$. Then, $\mathcal{FNN}_{d,\infty}^\rho$ is dense in $\mathcal{L}ip_\psi^{p,q}(\Lambda_T^\alpha)$, i.e. for every $F\in\mathcal{L}ip_\psi^{p,q}(\Lambda_T^\alpha)$ and $\varepsilon>0$ there exists some $\Phi\in\mathcal{FNN}_{d,\infty}^\rho$ such that*

$$\|F-\Phi\|_{\psi,p,q}<\varepsilon.$$

Proof. Since $\mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$ is defined as the $\|\cdot\|_{\psi,p,q}$-completion of $\mathcal{L}\mathrm{ip}_b^{p,q}(\Lambda_T^\alpha)$, we only need to approximate some $F\in\mathcal{L}\mathrm{ip}_b^{p,q}(\Lambda_T^\alpha)$ by an element of $\mathcal{FNN}_{d,\infty}^\rho$. Hereby, we observe that the vector space of cylindrical functionals

$$\mathcal{C}=\left\{\Lambda_T^\alpha\ni(t,Y)\mapsto\sum_{k=0}^{n}\mathbb{1}_{\{t>t_k\}}g_k(t-t_k,Y(t)-Y(t_k),Y(t_0),...,Y(t_k))\in\mathbb{R}\right\},\tag{4.3.2}$$

is dense in $\mathbb{C}_b^{p,q}(\Lambda_T^\alpha)$ with respect to $\|\cdot\|_{p,q}$, where $0 \leqslant t_0 \leqslant ... \leqslant t_n \leqslant T$ is a partition of $[0,T]$, and where $g_k \in C_b^\infty([0,T] \times \mathbb{R}^d \times \mathbb{R}^{(k+1)d})$ are smooth functions with $g_k(0,x,y_0,...,y_k) = g_k(t,0,y_0,...,y_k) = \frac{\partial^j g_k}{\partial t^j}(0,x,y_0,...,y_k) = \frac{\partial^{|\mathbf{m}|} g_k}{\partial x^{\mathbf{m}}}(t,0,y_0,...,y_k) = 0$, for all $j = 1,...,p$ and $1 \leqslant |\mathbf{m}| \leqslant q$. Since $\|\cdot\|_{\psi,p,q}$ induces a weaker topology on $\mathcal{L}\mathrm{ip}_b^{p,q}(\Lambda_T^\alpha) \subset \mathbb{C}_b^{p,q}(\Lambda_T^\alpha)$, there exists a cylindrical functional $G \in \mathcal{C}$ such that

$$\|F - G\|_{\psi,p,q} < \frac{\varepsilon}{3}. \tag{4.3.3}$$

Hereby, the cylindrical functional $G : \Lambda_T^\alpha \to \mathbb{R}$ can be considered as smooth map $G : X_0 \to \mathbb{R}$ defined on $X_0 = [0,T] \times \mathbb{R}^{(n+2)d}/\sim_n$ with equivalence relation given by $(t,x,y_0,...,y_n) \sim_n (s,w,z_0,...,z_n)$ if and only if $s = t$, $x = w$ and $y_k = z_k$, for all $k = 0,...,n_t$, with $n_t = \max\{k \in \{0,...,n\};\, t_k < t\}$. Hence, X_0 is a quotient manifold of $[0,T] \times \mathbb{R}^{(n+2)d}$ over $\mathbb{R}^{1+(n+2)d}$, where the horizontal and vertical derivatives can be identified with the partial derivatives in the first and second variable. In addition, we choose $Z = Y = \mathbb{R}$ as space for the hidden layer and output space, where $(\mathbb{R}, \psi_\mathbb{R})$ is a weighted space with $\psi_\mathbb{R}(z) = \exp(|z|)$. Hereby, we use the additive family between Λ_T^α and $\mathbb{R}$ given by $(t,Y) \mapsto H(t,Y) := \int_0^t \phi(s)^\top Y(s)ds + A^\top Y(t) + Bt + b$, where $\phi_i \in \mathcal{NN}_{1,\infty}^\rho$ is a conventional neural networks, and where $A \in \mathbb{R}^d$ and $B, b \in \mathbb{R}$, such that

$$\mathcal{H} = \left\{ h(t,x,y_0,...,y_n) = \sum_{l=0}^{n-1} \phi(t_l)^\top y_l(t \wedge t_{l+1} - t \wedge t_l) + A^\top x + Bt + b; \right.$$
$$\left. \phi \in \mathcal{NN}_{1,\infty}^\rho,\, A \in \mathbb{R}^d,\, B, b \in \mathbb{R} \right\}$$

is by construction a strong additive family between X_0 and $\mathbb{R}$. Since $\mathcal{NN}_{\mathbb{R},\mathbb{R}}^{\mathcal{G},\{\rho\},\mathcal{K}}$ is by Remark 3.50 $\tau_{\psi_\mathbb{R}}^{\max(p,q)}$-dense in $\mathcal{L}\mathrm{ip}_{\psi_\mathbb{R}}^{\max(p,q)}(\mathbb{R})$ and the topology induced by $\|\cdot\|_{\psi,p,q}$ is on the finite dimensional manifold X_0 weaker than $\tau_\psi^{\max(p,q)}$, there exists by Theorem 3.52 for the cylindrical functional $G : \Lambda_T^\alpha \to \mathbb{R}$ an abstract neural network of the form $\varphi(t,Y) = \sum_{k=1}^N \beta_k \rho(h_k(t,Y))$ satisfying

$$\|G - \varphi\|_{\psi,p,q} < \frac{\varepsilon}{3}. \tag{4.3.4}$$

Hereby, we observe that the non-anticipative functionals $h_k \in \mathcal{H}$ converge to their continuous versions $H_k(t,Y) := \int_0^t \phi_k(s)^\top Y(s)ds + A_k^\top Y(t) + B_k t + b_k$.

Indeed, by the mean value theorem there is $C_2 > 0$ with

$$\begin{aligned}
|H_k(t,Y) - h_k(t,Y)| &\leqslant \sum_{l=0}^{n-1} \left| \int_{t\wedge t_l}^{t\wedge t_{l+1}} (\varphi_k(s) - \varphi_k(t_l))^\top Y(s) ds \right| \\
&\leqslant C_2 h \|Y\|_{C^0([0,T];\mathbb{R}^d)} \leqslant C_2 h \psi(t,Y), \\
\left|\mathcal{D}^j H_k(t,Y) - \mathcal{D}^j h_k(t,Y)\right| &\leqslant \left| \phi_k^{(j)}(t)^\top Y(t) - \phi_k^{(j)}(t_{l_t})^\top Y(t_{l_t}) \right| \\
&\leqslant C_2 h \|Y\|_{C^0([0,T];\mathbb{R}^d)} \leqslant C_2 h \psi(t,Y),
\end{aligned}$$

for all $j = 1, ..., p$, where $l_t = \max\{l = 1, ..., n;\, t_l < t\}$. Since it also holds that $\partial_{\mathbf{m}} h_k(t,Y) = \partial_{\mathbf{m}} H_k(t,Y)$, for all $1 \leqslant |\mathbf{m}| \leqslant q$ and $(t,Y) \in \Lambda_T^\alpha$, we have $\|h_k - H_k\|_{\psi,p,q} < C_2 h$. Now, we use that the composition $\mathcal{L}\mathrm{ip}^{\max(p,q)}(X_0) \ni f \mapsto \rho \circ f \in \mathcal{L}\mathrm{ip}^{\max(p,q)}(X_0)$ is by [FK88, Proposition 4.4.15] continuous such that $\mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha) \ni h_k \mapsto \rho \circ h_k \in \mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$ is also continuous. Choosing $h > 0$ small enough and defining the functional neural network $\Phi \in \mathcal{FN}_{d,\infty}^\rho$ by $\Phi(t,Y) = \sum_{k=1}^N \beta_k \rho(H_k(t,Y))$, for $(t,Y) \in \Lambda_T^\alpha$, it follows that

$$\|\varphi - \Phi\|_{\psi,p,q} < \frac{\varepsilon}{3}. \tag{4.3.5}$$

Hence, by combining (4.3.3) and (4.3.4) with (4.3.5), we conclude that

$$\|F - \Phi\|_{\psi,p,q} \leqslant \|F - G\|_{\psi,p,q} + \|G - \varphi\|_{\psi,p,q} + \|\varphi - \Phi\|_{\psi,p,q} < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Since $\varepsilon > 0$ and $F \in \mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$ were chosen arbitrarily, it follows that $\mathcal{FN}_{d,\infty}^\rho$ is dense in $\mathcal{L}\mathrm{ip}_\psi^{p,q}(\Lambda_T^\alpha)$. □

Remark 4.7. *We observe that the space and time-dependent term of a functional neural network, i.e. the terms involving A and B in (4.2.1), are needed such that the corresponding additive family is strong in the sense of Definition 3.51. In this case, the Nachbin result in Theorem 3.49(i) ensures the approximation of the horizontal and vertical derivatives in $\mathcal{L}ip_\psi^{p,q}(\Lambda_T^\alpha)$, for $p, q \geqslant 1$. However, for universal approximation of functional neural networks in $\mathcal{L}ip_\psi^{0,0}(\Lambda_T^\alpha)$ without derivatives, these terms could be omitted.*

Remark 4.8. *The argument in Theorem 4.6 can be iterated to obtain universal approximation of deep functional neural networks described in Remark 4.4, such that universal approximation for functional neural networks with finitely many hidden layers follows.*

4.4 Examples of Learned Non-Anticipative Functionals

In this section, we illustrate in two particular examples how non-anticipative functionals can be learned with functional neural networks, where we consider the weighted space (Λ_T^α, ψ) described in Remark 4.5 with some fixed $\alpha < 1/2$. Since the sample paths of a one-dimensional Brownian motion $B = (B(t))_{t\in[0,T]}$ are a.s. continuous and $(1/2-\varepsilon)$-Hölder continuous, for all $\varepsilon > 0$, it follows thatthe stopped path (t, B^t) belongs a.s. to Λ_T^α. In order to learn the non-anticipative functionals, we use the package `Tensorflow` in `Python` to build and train the FNNs, with $N = 100$ number of neurons and conventional NNs φ_{ki} of size $L = (1, 50, 1)$, where we choose $T = 1$ and generate 10^5 sample paths of Brownian motions on $[0, 1]$, discretized over $n = 200$ points, which are split up into 80% for training and 20% for testing. The Adam algorithm (see [KB15]) is applied over 10^5 epochs with learning rate 10^{-4} and batchsize 500.

For the first example, we define $F : \Lambda_T^\alpha \to \mathbb{R}$ as running maximum $F(t, Y) = \sup_{s\in[0,t]} Y(s)$, which is not continuous with respect to d_∞. Following [Dup09], we consider an approximation in $\mathbb{C}^{0,0}(\Lambda_T^\alpha)$. For some $\varepsilon > 0$ and continuous function $g : \mathbb{R} \to \mathbb{R}$ satisfying $g(0) = 0$ and $g(2\varepsilon) = \varepsilon$, we define the approximation $F_{\varepsilon,g} : \Lambda_T^\alpha \to \mathbb{R}$ by

$$F_{\varepsilon,g}(t, Y) = \begin{cases} F(t, Y) - \varepsilon & 0 \leqslant Y(t) \leqslant F(t, Y) - 2\varepsilon \\ F(t, Y) - \varepsilon + g(Y(t) - F(t, Y) + 2\varepsilon) & F(t, Y) - 2\varepsilon < Y(t) \\ & \text{and } Y(t) \leqslant F(t, Y) \\ Y(t) & F(t, Y) < Y(t) \end{cases}$$

for $(t, Y) \in \Lambda_T^\alpha$, such that $F_{\varepsilon,g} \in \mathbb{C}^{0,0}(\Lambda_T^\alpha)$. In Figure 4.1, we depict the approximating running maximum $F_{\varepsilon,g} : \Lambda_T^\alpha \to \mathbb{R}$ with $g(x) = x/2$ and $\varepsilon = 0.001$ by a functional neural network $\Phi \in \mathcal{FN}_{d,\infty}^\rho$ in $\mathcal{L}\mathrm{ip}_\psi^{0,0}(\Lambda_T^\alpha)$.

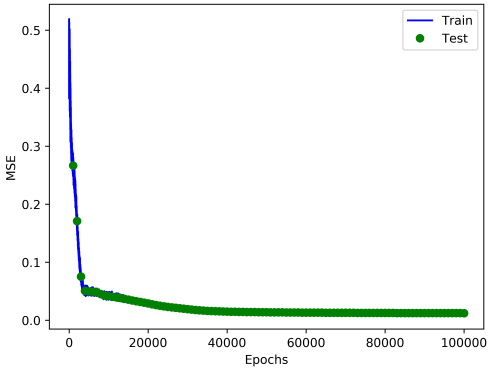
In the second example, we consider the non-anticipative functional $F : \Lambda_T^\alpha \to \mathbb{R}$ defined by

$$F(t, Y) = \frac{1}{2}\left(Y(t \wedge 1/2)^2 + Y(t)^2\right) + \int_0^t Y(s)^2 ds,$$

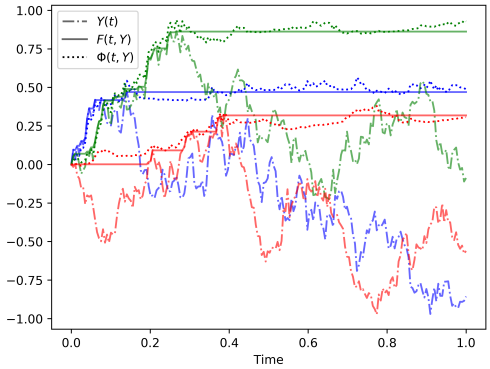
for $(t, Y) \in \Lambda_T^\alpha$, whose first horizontal and vertical derivative are given by

$$\mathcal{D}F(t, Y) = Y(t)^2, \qquad \nabla F(t, Y) = \begin{cases} 2Y(t) & t \leqslant 1/2 \\ Y(t) & t > 1/2 \end{cases}.$$

The results for the approximation in $\mathcal{L}\mathrm{ip}_\psi^{1,1}(\Lambda_T^\alpha)$ are displayed in Figure 4.2.

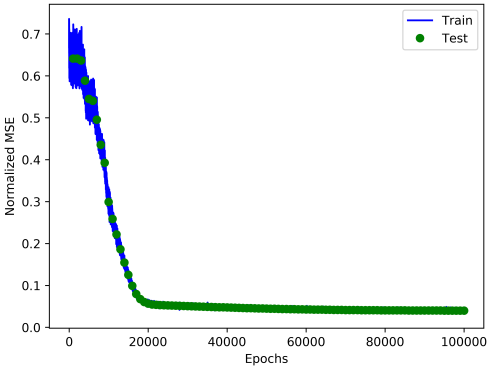


(a) Learning Performance

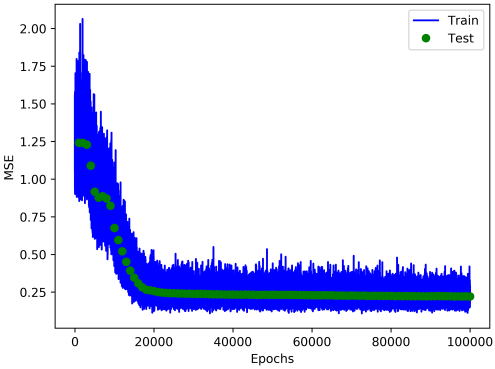


(b) $F_{\varepsilon,h}$ and Φ on three samples of test set

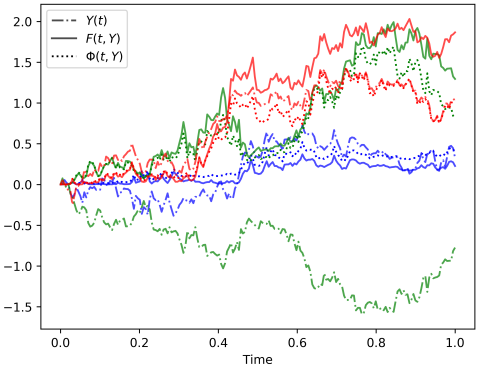
Figure 4.1: Learning the approximating running maximum $F_{\varepsilon,h} : \Lambda_T^\alpha \to \mathbb{R}$ by some FNN $\Phi : \Lambda_T^\alpha \to \mathbb{R}$ in the space $\mathcal{L}\mathrm{ip}_\psi^{0,0}(\Lambda_T^\alpha)$.



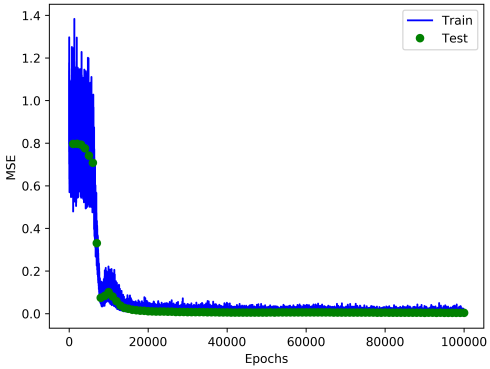
(a) Learning Performance



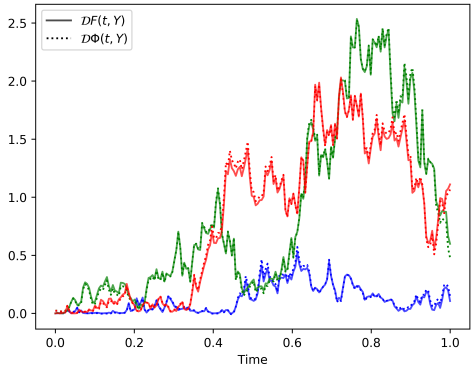
(b) Learning Performance for F



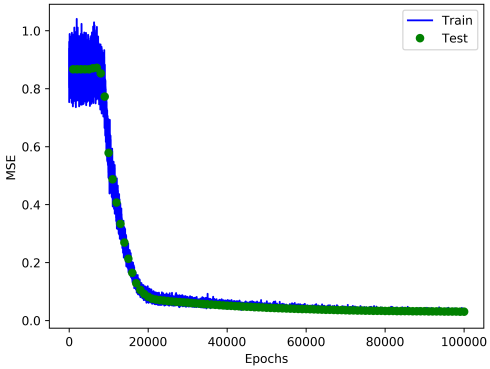
(c) F and Φ on three samples of test set



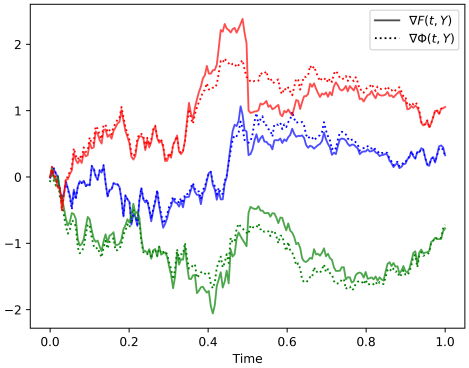
(d) Learning Performance for $\mathcal{D}F$



(e) $\mathcal{D}F$ and $\mathcal{D}\Phi$ on three samples of test set



(f) Learning Performance for ∇F



(g) ∇F and $\nabla \Phi$ on three samples of test set

Figure 4.2: Learning $F : \Lambda_T^\alpha \to \mathbb{R}$ by some FNN $\Phi : \Lambda_T^\alpha \to \mathbb{R}$ in the space $\mathcal{L}\mathrm{ip}_\psi^{1,1}(\Lambda_T^\alpha)$, where we normalize the mean squared error (MSE) for network training to account for the different magnitudes of F, $\mathcal{D}F$ and ∇F.

Part II

Learning Arbitrages in Stochastic Portfolio Theory

Chapter 5

Stochastic Portfolio Theory

Stochastic Portfolio Theory (SPT) is a novel stream in financial mathematics initiated by Robert Fernholz and summarized in his monograph [Fer02], which analyzes portfolio behaviour in equity markets. Opposed to classical portfolio theory introduced by Markowitz in [Mar52], SPT is descriptive and consistent with observable characteristics of portfolios in real equity markets, where the assumptions of the theory can be easily verified and hold true in actual equity markets. The main advantage of SPT is that no unobservable market characteristics such as the drift process of the stock prices need to be estimated. This is a new approach in contrast to the classical asset pricing theories, e.g. general equilibrium theory in [Arr53], capital asset pricing models in [Sha64] and [Mer69], as well as the market equilibrium characterized by the absence of arbitrage and existence of equivalent martingale measures in [HK79] and [DS94]. In this chapter we follow the notation and results of [Fer02].

5.1 Market and Portfolios

In this section, we introduce the financial market model consisting of $d \in \mathbb{N}$ assets with infinite time horizon, where the vector of stock prices is denoted by $S = (S_1(t), ..., S_d(t))_{t\in[0,\infty[}$. In many cases, S is modelled as a stochastic process defined on a filtered probability space $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{P})$ such that S is $\mathbb{F}$-adapted. In addition, we introduce the capitalization X defined for each stock $i = 1, ..., d$ as the price multiplied by the number of outstanding shares, i.e. $X_i(t) = n_i(t)S_i(t)$, for all $i = 1, ..., d$ and $t \in [0, T]$, where $n_i(t)$ denotes the number of outstanding shares of stock i at time t.

As in [Fer02, Definition 1.1], we assume that the vector of capitalizations

X follows a logarithmic Ito process, i.e.

$$\log X_i(t) = \log X_i(0) + \int_0^t \gamma_i(s)ds + \sum_{j=1}^m \int_0^t \xi_{ij}(s)dB_j(s), \tag{5.1.1}$$

for all $i = 1, ..., d$ and $t \in [0, \infty[$, where $B = (B_1, ..., B_m)$ is an m-dimensional Brownian motion, where the *growth rate process* $\gamma : \Omega \times [0, T] \to \mathbb{R}^d$ is measurable, $\mathbb{F}$-adapted and satisfies $\int_0^T \|\gamma(t)\|dt < \infty$, a.s., for all $T \in [0, \infty[$, and where the *volatility process* $\xi : [0, \infty[\to \mathbb{R}^{d\times d}$ is measurable, adapted and satisfies

$$\begin{aligned}
&\int_0^T \left(\xi_{i1}(t)^2 + ... + \xi_{id}(t)^2\right) dt < \infty,\\
&\lim_{t\to\infty} \frac{\left(\xi_{i1}(t)^2 + ... + \xi_{id}(t)^2\right)\log(\log(t))}{t} = 0,\\
&\xi_{i1}(t)^2 + ... + \xi_{id}(t)^2 > 0,
\end{aligned}$$

a.s., for all $i = 1, ..., d$, $T \in [0, \infty[$ resp. $t \in [0, \infty[$. Reformulating (5.1.1) in exponential form, we see that the capitalizations can be expressed as

$$X_i(t) = X_i(0) \exp\left(\int_0^t \gamma_i(s)ds + \sum_{j=1}^m \int_0^t \xi_{ij}(s)dB_j(s)\right),$$

for all $i = 1, ..., d$ and $t \in [0, \infty[$, such that $X_i(t) > 0$ if the initial capitalization satisfies $X_i(0) > 0$.

From the volatilities $\xi = (\xi_{ij})_{ij}$, we define the *covariance process* $\sigma = (\sigma_{ij})_{ij} : \Omega \times [0, \infty[\to \mathbb{R}^{d\times d}$ by $\sigma(t) = \xi(t)\xi(t)^\top$, for $t \in [0, \infty[$, which is by definition a positive semidefinite process, i.e. $x^\top \sigma(t)x \geqslant 0$, a.s., for all $x \in \mathbb{R}^d$ and $t \in [0, \infty[$. In this case, we observe that

$$\langle \log X_i, \log X_j \rangle_t = \int_0^t \sigma_{ij}(s)ds = \sum_{k=1}^d \int_0^t \xi_{ik}(s)\xi_{kj}(s)ds,$$

a.s., for all $i, j = 1, ..., d$ and $t \in [0, \infty[$.

Definition 5.1. *The stock price process X is called a* market *if $\sigma(t)$ is a.s. non-singular, for all $t \in [0, \infty[$. In this case, a market X is* non-degenerate *if there exists a number $\beta > 0$ such that*

$$x^\top \sigma(t) x \geqslant \beta \|x\|^2, \tag{ND_β}$$

a.s., for all $t \in [0, \infty[$. A market X has bounded variance *if there exists a number $\kappa > 0$ such that*

$$x^\top \sigma(t) x \leqslant \kappa \|x\|^2, \tag{BV_κ}$$

a.s., for all $t \in [0, \infty[$.

The non-degeneracy and bounded variance condition pertain the most in real markets, reflecting that the variance cannot become arbitrarily large or arbitrarily small. From now on, we shall always assume that the number of Brownian motions is equal to the dimension of the market, i.e. $m = d$, and that σ is assumed to be non-singular as in Definition 5.1. Next, we introduce portfolios as investment strategies.

Definition 5.2. *A* portfolio *π is a measurable, $\mathbb{F}$-adapted, $\mathbb{R}^d$-valued process $\pi = (\pi_1(t), ..., \pi_d(t))_{t \in [0,T]}$ such that π is a.s. bounded and it holds that $\pi_1(t) + ... + \pi_d(t) = 1$, a.s., for all $t \in [0, \infty[$. In this case, a portfolio π is called* long-only *if $\pi_i(t) \geqslant 0$, for all $i = 1, ..., d$ and $t \in [0, \infty[$.*

If π is a portfolio, $\pi_i(t)$ denotes the *fraction of wealth* invested into the ith stock, such that $\pi_1, ..., \pi_d$ are called the *portfolio weights*. Suppose that $Z_\pi(t) > 0$ represents the value of the portfolio π at time t, then the amount invested in stock i is equal to $\pi_i(t) Z_\pi(t)$. Hence, the change of the portfolio value induced by a change in stock i is given by $\pi_i(t) Z_\pi(t) \frac{dX_i(t)}{X_i(t)}$ such that the total change in the portfolio value is equal to

$$dZ_\pi(t) = \sum_{i=1}^{d} \pi_i(t) Z_\pi(t) \frac{dX_i(t)}{X_i(t)},$$

or equivalently

$$\frac{dZ_\pi(t)}{Z_\pi(t)} = \sum_{i=1}^{d} \pi_i(t) \frac{dX_i(t)}{X_i(t)}, \tag{5.1.2}$$

for $t \in [0, \infty[$. This equation expresses the instantaneous return of the portfolio value $\frac{dZ_\pi(t)}{Z_\pi(t)}$ as weighted sum of the instantaneous return of the stocks $\frac{dX_i(t)}{X_i(t)}$ with respect to the portfolio weights.

Proposition 5.3 ([Fer02, Proposition 1.1.5]). *Let π be a portfolio in the market X. Then, it holds that*

$$\log Z_\pi(t) = \log Z_\pi(0) + \int_0^t \gamma_\pi(s)ds + \sum_{i,j=1}^d \int_0^t \pi_i(s)\xi_{ij}(s)dB_j(s),$$

a.s., for all $t \in [0,\infty[$, where the portfolio growth rate γ_π *is given by*

$$\gamma_\pi(t) = \sum_{i=1}^d \pi_i(t)\gamma_i(t) + \underbrace{\frac{1}{2}\left(\sum_{i=1}^d \pi_i(t)\sigma_{ii}(t) - \sigma_{\pi\pi}(t)\right)}_{=:\gamma_\pi^\star(t)},$$

where $\sigma_{\pi\pi}$ denotes the portfolio variance *given by $\sigma_{\pi\pi}(t) = \pi(t)^\top\sigma(t)\pi(t)$, for $t \in [0,\infty[$, such that $\langle \log Z_\pi\rangle_t = \int_0^t \sigma_{\pi\pi}(s)ds$, a.s., for all $t \in [0,\infty[$, and where γ_π^* represents the* portfolio excess growth rate.

Corollary 5.4 ([Fer02, Corollary 1.1.6]). *Let π be a portfolio with value process Z_π. Then, it follows that*

$$\log Z_\pi(t) = \log Z_\pi(0) + \sum_{i=1}^d \int_0^t \pi_i(s)d\log X_i(s) + \int_0^t \gamma_\pi^*(s)ds,$$

a.s., for all $t \in [0,\infty[$.

These expressions of the portfolio value can be used to measure the performance of stocks or portfolios relative to another portfolio, which is frequently of interest for example with respect to the S&P500. A natural benchmark is the market portfolio consisting of all shares of all stocks in the market.

Definition 5.5. *Let X_i be a stock in the market X and η be a portfolio. The* relative return *of X_i with respect to η is given by $(\log(X_i(t)/Z_\eta(t)))_{t\in[0,\infty[}$.*

Defining $\sigma_{i\eta}$ by $\sigma_{i\eta}(t) = \sum_{j=1}^d \sigma_{ij}(t)\eta_j(t)$, for $i = 1,...,d$ and $t \in [0,\infty[$, the *relative covariance process* $\tau^\eta = (\tau_{ij}^\eta(t))_{t\in[0,\infty[,\, i,j=1,...,d}$ is given by

$$\tau_{ij}^\eta = \sigma_{ij}(t) - \sigma_{i\eta}(t) - \sigma_{j\eta}(t) + \sigma_{\eta\eta}(t),$$

for $i,j = 1,...,d$ and $t \in [0,\infty[$. In this case, we observe that we have $\langle \log(X_i/Z_\eta), \log(X_j/Z_\eta)\rangle_t = \int_0^t \tau_{ij}^\eta(s)ds$, a.s., for all $i,j = 1,...,d$ and $t \in [0,\infty[$.

Lemma 5.6 ([Fer02, Lemma 1.2.2])**.** *For a portfolio η, $\tau^\eta(t)$ is a.s. positive semidefinite with rank $d-1$, for all $t \in [0,\infty[$. In addition, the null space of $\tau^\eta(t)$ is spanned by $\eta(t)$.*

Next, we introduce the market portfolio, which we will use often as benchmark in the following.

Definition 5.7. *The portfolio μ with weights defined by*

$$\mu_i(t) = \frac{X_i(t)}{X_1(t) + ... + X_d(t)} \quad \in \quad \Delta^d = \left\{x \in]0,1[^d;\, x_1 + ... + x_d = 1\right\}, \tag{5.1.3}$$

for $i = 1, ..., d$ and $t \in [0,\infty[$, is called the market portfolio*, and the weights are called the* market weights*.*

For the market portfolio, it holds that $0 < \mu_i(t) < 1$, for all $i = 1, ..., d$ and $t \in [0,\infty[$, such that its weights take values in the *open simplex* Δ^d defined on the right-hand side of (5.1.3). In addition, the corresponding value process Z_μ is given by $Z_\mu(t) = X_1(t) + ... + X_d(t)$, for all $t \in [0,\infty[$, because Z_μ satisfies (5.1.2) with weights $\mu_i(t)$ defined in (5.1.3). Hence, $Z_\mu(t)$ represents the capitalizations of all stocks in the market. Moreover, the logarithmic market weights $\log \mu_i(t) = \log(X_i(t)/Z_\mu(t))$ denotes the relative return of the ith stock with respect to the market portfolio.

We denote the relative covariance of the market portfolio by τ instead of τ^μ, where we observe that

$$\langle \log \mu_i, \log \mu_j \rangle_t = \left\langle \log\left(\frac{X_i}{Z_\mu}\right), \log\left(\frac{X_j}{Z_\mu}\right) \right\rangle_t = \int_0^t \tau_{ij}(s) ds,$$

a.s., for all $i, j = 1, ..., d$ and $t \in [0,\infty[$. Furthermore, by applying Ito's formula, we conclude that

$$\langle \mu_i, \mu_j \rangle_t = \int_0^t \mu_i(s)\mu_j(s) d\langle \log \mu_i, \log \mu_j \rangle_s = \int_0^t \mu_i(s)\mu_j(s)\tau_{ij}(s) ds,$$

a.s., for all $i, j = 1, ..., d$ and $t \in [0,\infty[$. In addition, we introduce the relative return of a portfolio with respect to another portfolio.

Definition 5.8. *Let π and η be two portfolios. Then, the* relative return *of π with respect to η is defined as* $\log\left(Z_\pi(t)/Z_\eta(t)\right)$*, for $t \in [0,\infty[$.*

If $\eta = \mu$ is the market portfolio, we obtain by Corollary 5.4 that the relative

return satisfies

$$\log\left(\frac{Z_\pi(t)}{Z_\mu(t)}\right) = \log\left(\frac{Z_\pi(0)}{Z_\mu(0)}\right) + \sum_{i=1}^{d}\int_0^t \pi_i(s)d\log\mu_i(s) + \int_0^t \gamma_\pi^*(s)ds,$$

a.s., for all $t \in [0,\infty[$. In addition, it follows by [Fer02, Proposition 1.3.1] that the growth rate of the portfolio determines its long term behaviour because we have

$$\lim_{T\to\infty}\frac{1}{T}\left(\log Z_\pi(T) - \int_0^T \gamma_\pi(t)dt\right) = 0, \quad \text{a.s.}$$

Using this and the other ingredients of the market introduced in this section, we consider two different portfolio optimization problem.

Example 5.9 ([Mar52])**.** *The classical Markowitz portfolio optimization problem in [Mar52] is given by minimizing the portfolio variance under a drift constraint, i.e.*

$$\min_{\substack{\pi \text{ long-only}\\ \text{portfolio}}} \sigma_{\pi\pi}(t) \qquad \text{subject to} \qquad \sum_{i=1}^{d}\left(\gamma_i(t) + \frac{1}{2}\sigma_{ii}(t)\right) \geqslant \alpha_0,$$

for fixed $t > 0$ and given $\alpha_0 \in \mathbb{R}$, which can be solved with conventional quadratic programming methods.

Example 5.10 ([Fer02, Example 1.3.8])**.** *We consider the following optimization problem, which does not involve the estimation of the drift process γ_i. Let η be a benchmark portfolio and assume that all stocks have the same drift process γ, then the growth rate of η is given by $\gamma_\eta(t) = \gamma(t) + \frac{1}{2}\sum_{i=1}^{d}\eta_i(t)\tau_{ii}^\eta(t)$, a.s., for all $t \in [0,\infty[$. If the tracking error of π with respect to η is small enough, then minimizing*

$$\min_{\substack{\pi \text{ long-only}\\ \text{portfolio}}} \pi(t)^\top\tau^\eta(t)\pi(t) \quad \text{subject to} \quad \frac{1}{2}\sum_{i=1}^{d}\pi_i(t)\tau_{ii}^\eta(t) \geqslant \gamma_0 + \frac{1}{2}\sum_{i=1}^{d}\eta_i(t)\tau_{ii}^\eta(t),$$

with fixed $t > 0$ and some $\gamma_0 > 0$, yields a portfolio with growth rate about γ_0 greater than the benchmark's growth rate, where conventional quadratic programming can be applied to solve the minimization problem.

5.2 Market Behaviour and Diversity

Stochastic Portfolio Theory was introduced as a new concept in financial mathematics to study the long-term behaviour of portfolios under reasonable assumptions on the market. Such assumptions were the non-degeneracy condition in (ND_β) and the bounded variance condition in (BV_κ) introduced in Definition 5.1, whereas we introduce two other conditions called "coherence" and "diversity" in this section.

Definition 5.11. *The market X is* coherent *if* $\lim_{t\to\infty} \log(\mu_i(t))/t = 0$, *a.s., for all* $i = 1, ..., d$.

Since $\log \mu_i(t) < 0$, the market is coherent if none of the stocks decreases too rapidly. In a coherent market, [Fer02, Proposition 2.1.2] shows that the difference between the time-average growth rates of two distinct stocks must be zero. For example, if all stocks have the same growth rate, then it follows by [Fer02, Corollary 2.1.3] that the market is coherent. Assuming in addition that the market is non-degenerate, we find the following (relative) arbitrage opportunity.

Proposition 5.12 ([Fer02, Proposition 2.1.9])**.** *Assume that the market X is non-degenerate and coherent, and let π be the constant-weighted portfolio given by $\pi_i(t) = p_i$, for all $i = 1, ..., d$, with $p_1 + ... + p_d = 1$ and $p_1, ..., p_d \geqslant 0$ as well as $p_{i_1}, p_{i_2} > 0$, for at least two distinct indices $i_1, i_2 \in \{1, ..., d\}$. Then, it follows that*

$$\liminf_{T\to\infty} \frac{1}{T} \log\left(\frac{Z_\pi(T)}{Z_\mu(T)}\right) > 0, \quad a.s.$$

Besides the coherence condition, market diversity plays a crucial role in real equity markets. Diversity means that none of the assets can become arbitrarily large and account for the entire market capitalization.

Definition 5.13. *A market X is called* diverse *if there exists a number $\delta > 0$ such that for every $t \in [0, \infty[$, it holds that*

$$\max_{i=1,...,d} \mu_i(t) \leqslant 1 - \delta. \qquad (DV_\delta)$$

A market X is called weakly diverse *if there exists a number $\delta > 0$ such that $\frac{1}{T}\int_0^T \max_{i=1,...,d} \mu_i(t)dt \leqslant 1 - \delta$, for all $T \in [0, \infty[$.*

Even though the diversity condition is a fairly weak assumption on the market, it has strong mathematical consequences. In the following, we give

the relation of diversity to the excess growth rate of the market portfolio and summarize the other results in [Fer02, Section 2.2] subsequently.

Proposition 5.14 ([Fer02, Proposition 2.2.2])**.** *Assume that the market X is non-degenerate and diverse. Then, there exists some $\delta > 0$ with*

$$\gamma_\mu^*(t) \geqslant \delta, \tag{5.2.1}$$

a.s., for all $t \in [0, \infty[$. Conversely, if X has bounded variance and there exists some $\delta > 0$ such that (5.2.1) *holds true, then X is diverse.*

Moreover, if all the stocks in the market X have the same growth rate, it follows by [Fer02, Proposition 2.2.3] that $\lim_{T\to\infty} \frac{1}{T} \int_0^T \gamma_\mu^*(t)dt = 0$, a.s., which implies that if X is additionally non-degenerate that X is not diverse (see [Fer02, Corollary 2.2.4]). Finally, we introduce the entropy function $S(x) = -\sum_{i=1}^d x_i \log x_i$, which can be seen as a measure of diversity, determining the uniformity of a probability measure, for which the following equivalent characterization of diversity holds true.

Proposition 5.15 ([Fer02, Proposition 2.3.2])**.** *The market X is diverse if and only if there exists some $\varepsilon > 0$ such that*

$$S(\mu(t)) \geqslant \varepsilon,$$

a.s., for all $t \in [0, \infty[$.

With the help of [Fer02, Theorem 2.3.4], we can look at the value process of the particular *entropy-weighted portfolio* π, whose weights are defined as the logarithmic gradient of the entropy function, i.e.

$$\pi_i(t) = \frac{-\mu_i(t) \log \mu_i(t)}{S(\mu(t))},$$

for $i = 1, ..., d$ and $t \in [0, \infty[$. This portfolio can be used to outperform the market portfolio in the long-run.

Corollary 5.16 ([Fer02, Corollary 2.3.5])**.** *Let π be the entropy-weighted portfolio in a non-degenerate and diverse market. Then, for sufficiently large $T > 0$, it holds that*

$$\frac{Z_\pi(T)}{Z_\pi(0)} > \frac{Z_\mu(T)}{Z_\mu(0)}, \quad \text{a.s.}$$

The basic arguments used for Corollary 5.16 are the expression of the relative return for the entropy-weighted portfolio π with respect to the market portfolio μ in terms of the excess growth rate γ^*_μ given in [Fer02, Theorem 2.3.4] as well as Proposition 5.14 and the equivalent characterization in Proposition 5.15.

5.3 Functionally Generated Portfolios

The main ingredient of Stochastic Portfolio Theory (SPT) is to construct investment strategies, whose relative performace with respect to the market portfolio is captured in a simple equation. One particular example of such a portfolio was presented in the previous section by the entropy-weighted portfolio, which is constructed with the help of the entropy function $S(x) = -\sum_{i=1}^d x_i \log x_i$ measuring the diversity of the market. In this section, we consider more general functions defined on the open simplex Δ^d whose logarithmic gradient can be used to construct a portfolio.

Definition 5.17. *Let $S : \Delta^d \to [0, \infty[$ be a strictly positive continuous function and let π be a portfolio. Then, $S : \Delta^d \to [0, \infty[$* generates *the portfolio π if there exists a measurable process of finite variation Θ such that*

$$\log\left(\frac{Z_\pi(t)}{Z_\mu(t)}\right) = \log S(\mu(t)) + \Theta(t), \tag{5.3.1}$$

a.s., for all $t \in [0, \infty[$. Then, $S : \Delta^d \to [0, \infty[$ is called the generating function *of π and the process Θ is called the* drift process.

Since Θ is of finite variation and the relative return $t \mapsto \log(Z_\pi(t)/Z_\mu(t))$ is a semimartingale, $t \mapsto \log S(\mu(t))$ is also a semimartingale. Moreover, (5.3.1) provides a decomposition of the relative return into a part involving the generating function and the drift part. Hereby, the generating function part usually determines the short-term behaviour and the drift component the long-term behaviour of the portfolio, which is in particular the case if $\log S$ is bounded on Δ^d. In the following, we denote the partial derivatives of a function $f : U \to \mathbb{R}$ with $U \subset \mathbb{R}^d$ by $D_i f = \frac{\partial}{\partial x_i} f$ and $D_{ij} f = \frac{\partial^2}{\partial x_i \partial x_j} f$, for all $i, j = 1, ..., d$.

Theorem 5.18 ([Fer02, Theorem 3.1.5]). *Let $S : U \to [0, \infty[$ be a twice differentiable function defined on an open neighborhood U of Δ^d such that $x \mapsto x_i D_i \log S(x)$ is bounded on Δ^d. Then, $S : U \to [0, \infty[$ generates the portfolio π with weights given by*

$$\pi_i(t) = \left(D_i \log S(\mu(t)) + 1 - \sum_{j=1}^{d} \mu_j(t) D_j \log S(\mu(t)) \right) \mu_i(t), \tag{5.3.2}$$

for all $i = 1, ..., d$ and $t \in [0, T]$, and the corresponding drift process Θ is given by

$$\Theta(t) = \sum_{i,j=1}^{d} \int_0^t \frac{-1}{2S(\mu(s))} D_{ij} S(\mu(s)) \mu_i(s) \mu_j(s) \tau_{ij}(s) ds, \tag{5.3.3}$$

for all $t \in [0, T]$.

The particular advantage of Theorem 5.18 is that once an investor has chosen her investment map $S : U \to [0, \infty[$, the investment strategy in terms of portfolio weights and drift process is fully determined by (5.3.2) and (5.3.3). Moreover, the covariance structure of the market only needs to be estimated if one wants to know the long-term behaviour of the portfolio expressed by the drift process in (5.3.3). In the following, we present particular examples of generating functions.

Example 5.19 ([Fer02, Example 3.1.6]). *The following are examples of generating functions.*

(i) The constant function $S(x) = 1$ generates the market portfolio μ with $\Theta(t) = 0$.

(ii) The linear function $S(x) = w^\top x$, with $w \in \mathbb{R}^d$, generates the buy-and-hold portfolio holding w_i shares of the ith stock, where $\Theta(t) = 0$.

(iii) The smooth Gini function $S(x) = 1 - \frac{1}{2} \sum_{i=1}^{d} x_i^2$ generates the portfolio with weights

$$\pi_i(t) = \left(\frac{2 - \mu_i(t)}{S(\mu(t))} - 1 \right) \mu_i(t),$$

where $\Theta(t) = \sum_{i=1}^{d} \int_0^t \frac{1}{2S(\mu(s))} \mu_i(s)^2 \tau_{ii}(s) ds$.

Example 5.19 (continued)**.**

(iv) The entropy generating function $S(x) = -\sum_{i=1}^{d} x_i \log x_i$ *generates the entropy-weighted portfolio*

$$\pi_i(t) = -\frac{\mu_i(t)\log\mu_i(t)}{S(\mu(t))},$$

where $\Theta(t) = \int_0^t \frac{\gamma_\mu^*(s)}{S(\mu(s))} ds$.

(v) The geometric function $S(x) = \prod_{i=1}^{d} x_i^{p_i}$*, with* $p_1 + ... + p_d = 1$*, generates the constant-weighted portfolio with* $\pi_i(t) = p_i$ *and* $\Theta(t) = \int_0^t \gamma_\pi^*(s) ds$*. A special case is given by the choice of* $p_i = \frac{1}{d}$*, which represents the equally-weighted portfolio.*

(vi) The diversity function $S(x) = \left(\sum_{i=1}^{d} x_i^p\right)^{1/p}$*, for some* $p \neq 0$*, generates the diversity-weighted portfolio*

$$\pi_i(t) = \frac{\mu_i(t)^p}{S(\mu(t))^p},$$

where $\Theta(t) = (1-p)\int_0^t \gamma_\pi^*(s) ds$.

In addition, different generating functions can be aggregated to obtain a hybrid generating function. Let $S_1, ..., S_m : U \to [0, \infty[$ generate the portfolios π_1,...,π_m and let $p_1, ..., p_m \in \mathbb{R}$ be constants with $p_1 + ... + p_m = 1$. Then, the hybrid generating function $S = S_1^{p_1} \cdots S_m^{p_m}$ generates the portfolio with weights

$$\pi_i(t) = p_1\pi_{1i}(t) + ... + p_m\pi_{mi}(t),$$

for all $i = 1, ..., d$ and $t \in [0, \infty[$, where π_{ki} denotes the ith portfolio weight of the portfolio π_k. Let us also mention that not every portfolio is functionally generated, for which [Fer02, Proposition 3.1.11] gives sufficient conditions, where the components of the portfolio need to be a gradient of a real-valued function defined on an open neighborhood U of Δ^d.

In the classical theory of financial markets, the no-arbitrage hypothesis is usually assumed, stating that it is not possible to invest into the market with no initial capital and end up with non-negative terminal value having stricitly positive value for some positive probability. However, using Theorem 5.18 for portfolio generation in SPT, we can construct somehow unrealistic portfolios.

Example 5.20 ([Fer02, Example 3.3.3]). *Let X be a market satisfying (ND_β), which is weakly diverse with constant $\delta > 0$. Let $S : U \to [0, \infty[$ be the smooth Gini function in Example 5.19(iii), which generates the portfolio π. Then, π is admissible in the sense that there exist some numbers $c, M > 0$ such that*

$$\pi_i(t) \geqslant 0, \qquad \frac{Z_\pi(t)}{Z_\pi(0)} \geqslant c\frac{Z_\mu(t)}{Z_\mu(0)}, \qquad \frac{\pi_i(t)}{\mu_i(t)} \leqslant M,$$

a.s., for all $i = 1, ..., d$ and $t \in [0, \infty[$. Moreover, π strictly dominates the market portfolio μ in the sense that

$$\frac{Z_\pi(T)}{Z_\pi(0)} > \frac{Z_\mu(T)}{Z_\mu(0)}, \quad a.s.$$

which means that π outperforms the market portfolio over the time horizon $[0, T]$. However, the terminal time needs to satisfy $T > \frac{2d\log(2)}{\varepsilon\delta^2}$ such that one has to wait very long for the arbitrage. From this, we conclude that short-term arbitrages might be excluded in non-degenerate and weakly diverse markets.

Looking at functionally generated portfolios characterized in Theorem 5.18, we observe that the drift process depends on the Hessian of the generating function with a negative sign, which means that for an increasing drift process, the Hessian $D^2S = (D_{ij}S(x))_{i,j=1,...,d}$ must be negative semidefinite, for all $x \in \Delta^d$. This is satisfied if the generating function $S : \Delta^d \to [0, \infty[$ is *strictly concave*, i.e.

$$\lambda S(x) + (1 - \lambda)S(y) < S(\lambda x + (1 - \lambda)y),$$

for all $\lambda \in [0, 1]$ and $x, y \in \Delta^d$. In addition, a function $S : \Delta^d \to [0, \infty[$ is called *symmetric* if

$$S(x_1, ..., x_d) = S(x_{\sigma(1)}, ..., x_{\sigma(d)}),$$

for all $x \in \Delta^d$ and permutations $\sigma : \{1, ..., d\} \to \{1, ..., d\}$ of the indices $\{1, ..., d\}$. A function satisfying these two conditions is called a *measure of diversity*.

Proposition 5.21 ([Fer02, Proposition 3.4.2]). *Let $S : U \to [0, \infty[$ be symmetric and strictly concave, which generates the portfolio π with drift process Θ. Then, $t \mapsto \Theta(t)$ is a.s. non-decreasing and $\mu_i(t) \geqslant \mu_j(t)$ implies $\pi_j(t)/\mu_j(t) \geqslant \pi_i(t)/\mu_i(t)$, a.s., for all $t \in [0, \infty[$.*

Typical examples of measures of diversity are the entropy function, the geometric function and the diversity function in Example 5.19(iv)-(vi). Altogether, functionally generated portfolios are a natural generalization of the entropy-weighted portfolio introduced in the previous section.

5.4 Path-Dependent Setting

A closer look at functionally generated portfolios described in Theorem 5.18 reveals that such portfolios can only depend on the current value of the market weights, and therefore cannot take into account the history of the market. For this reason, [SSV18] introduced the so-called "path-dependent setting" of SPT, where a portfolio is generated by a path-dependent functional, which can depend on the previous values of the market portfolio. This approach relies on functional Ito calculus developed by [Dup09], [CF10] and [CF13], which generalizes Föllmer's pathwise Ito calculus to path-dependent functionals. For this section, we follow along the lines of [SSV18, Section 2.2] and fix some finite time horizon $T > 0$.

Definition 5.22. *We say that* $\pi : \Lambda_T \to \mathbb{R}^d$ *is a* functional portfolio *if it is an admissible functional integrand for* $\log X$ *on* $\mathbb{R}^d$ *such that* $\pi_1(t, \log X) + ... + \pi_d(t, \log X) = 1$*, for all* $t \in [0, T]$*, and* $t \mapsto \int_0^t \pi(s, \log X) d \log X(s)$ *admits along* $(\mathbb{T}_n)_{n\in\mathbb{N}}$ *the quadratic variation*

$$\left\langle \int_0^{\cdot} \pi(s, \log X) d \log X(s) \right\rangle_t = \sum_{i,j=1}^{d} \int_0^t \pi_i(s, \log X) \sigma_{ij}(s) \pi_j(s, \log X) ds, \tag{5.4.1}$$

for all $t \in [0, T]$.

Remark 5.23. *While* (5.4.1) *is always satisfied in the case without path-dependence, it is in general not fulfilled in this path-dependent setting, where [AC17, Theorem 2.1] gives sufficient conditions.*

Remark 5.24. *Choosing* $F(t, Y) = \log(e^{Y_1(t)} + ... + e^{Y_d(t)})$*, it follows that* $\nabla F(t, \log X) = \mu(t)$*, such that the market portfolio* μ *is a functional portfolio. Hereby,* (5.4.1) *holds true by [AC17, Theorem 2.1] and using that* $F \in \mathbb{C}^{1,2}(\Lambda_T)$ *is boundedness-preserving and uniformly Lipschitz continuous in* Y *with boundedness-preserving first vertical derivative* $\nabla F \in \mathbb{C}^{1,1}(\Lambda_T)$*.*

We now define the value process of a functional portfolio $\pi : \Lambda_T \to \mathbb{R}^d$. Inspired by the setting without path-dependence described in Proposition 5.3, we define the value of the portfolio $\pi : \Lambda_T \to \mathbb{R}^d$ as a process $Z_\pi = (Z_\pi(t))_{t\in[0,T]}$

given by

$$\begin{aligned}\log Z_\pi(t) = \log Z_\pi(0) &+ \sum_{i=1}^{d} \int_0^t \pi_i(s, \log X) d \log X_i(s) \\ &+ \frac{1}{2} \sum_{i=1}^{d} \int_0^t \pi_i(s, \log X) \sigma_{ii}(s) ds \\ &- \frac{1}{2} \sum_{i,j=1}^{d} \int_0^t \pi_i(s, \log X) \pi_j(s, \log X) \sigma_{ij}(s) ds,\end{aligned}$$

a.s., for all $t \in [0,T]$, where $Z_\pi(0)$ denotes the initial endowment of π. In addition, a non-anticipative functional $F : \Lambda_T \to [0, \infty[$ *generates* the functional portfolio π if there exists a path of finite variation $\Theta : [0,T] \to \mathbb{R}$ such that

$$\log \left(\frac{Z_\pi(t)}{Z_\mu(t)} \right) = \log F(t, \mu) + \Theta(t),$$

a.s., for all $t \in [0,T]$. We recall the main result of [SSV18], which describes the weights of a portfolio generated by the vertical derivatives of the smooth non-anticipative functional $F \in \mathbb{C}^{1,2}(\Lambda_T)$, where a non-anticipative functional $F : \Lambda_T \to \mathbb{R}$ is called *locally path-continuous* if for every $\varepsilon > 0$ and $(t, Y) \in [0,T] \times C^0([0,T])$ there exists some $\delta > 0$ such that $|F(u,Y) - F(u,Z)| < \varepsilon$, for all $(u, Z) \in [0,T] \times C^0([0,T])$ with $\|Y - Z\|_{C^0([0,T])} < \delta$ and $|t - u| < \delta$.

Theorem 5.25 ([SSV18, Theorem 2.16])**.** *Let $F \in \mathbb{C}^{1,2}(\Lambda_T)$ be a strictly positive non-anticipative functional, which is boundedness-preserving and locally path-continuous. Then, $F : \Lambda_T \to [0, \infty[$ generates the functional portfolio $\pi : \Lambda_T \to \mathbb{R}^d$ with weights given by*

$$\pi_i(t, \mu) = \left(\partial_i \log F(t, \mu) + 1 - \sum_{j=1}^{d} \mu_j(t) \partial_j \log F(t, \mu) \right) \mu_i(t),$$

a.s., for all $i = 1, ..., d$ and $t \in [0,T]$, such that

$$\Theta(t) = - \sum_{i,j=1}^{d} \int_0^t \frac{1}{2F(s,\mu)} \partial_{ij} F(s, \mu) \mu_i(s) \mu_j(s) \tau_{ij}(s) ds - \int_0^t \frac{\mathcal{D}F(s,\mu)}{F(s,\mu)} ds,$$

a.s., for all $t \in [0,T]$.

Of course, every functionally generated portfolio described in Example 5.19 can be transferred to the path-dependent setting by defining $F(t, \mu) = S(\mu(t))$,

for $(t, \mu) \in \Lambda_T$. More involved examples of path-dependent portfolios have been introduced in [SSV18, Section 3], for example the functional $F \in \mathbb{C}^{1,2}(\Lambda_T)$ defined by

$$F(t, \mu) = \varphi\left(\lambda\mu(t) + \frac{1-\lambda}{\theta}\int_{t-\theta}^{t} \mu(0 \wedge s)ds\right),$$

for $t \in [0, T]$ and some twice differentiable function $\varphi : \Delta^d \to [0, \infty[$, taking into account the past values of the market portfolio over the time horizon $[t - \theta, \theta]$. We refer to [SSV18, Section 3] for a detailed discussion of this class of generating functionals. In addition, we index sometimes the portfolio π^F by its generating functional $F : \Lambda_T \to [0, \infty[$, similarly as in the case without path-dependence.

Chapter 6

Pathwise Stochastic Portfolio Theory

In this chapter, we do not assume a priori any financial market model, nor do we restrict ourselves to certain stochastic processes modelling the stock prices. Instead, we consider a model-free approach, where the stock prices follow a deterministic path in $\mathbb{R}^d_+ := [0, \infty[^d$, for which we apply Föllmer's pathwise Ito calculus [Föl81] to define stochastic integration in a probability-free way. In this model-free setting, the notion of functionally generated portfolios within the framework SPT were obtained by [PW14] in discrete time and by [SSV18] in continuous time, where the latter allows for more general functionally-generated portfolios, whose generating functional can depend on time as well as the path of the market weights or other continuous trajectories of bounded variation. This is achieved by applying the generalized functional Ito calculus developed by [Dup09], [CF10] and [CF13], which generalizes Föllmer's Ito calculus to path-dependent functionals.

6.1 Market Weights and Portfolios

In the following, we assume that the financial market model consists of $d \in \mathbb{N}$ assets with finite time horizon $T > 0$, whose price vector is given by $S = (S_1(t), ..., S_d(t))_{t\in[0,T]}$. In many cases, such as in the previous chapter, S is modelled by a stochastic process defined on a probability space, whereas we consider in this chapter the model-free approach, where S is a deterministic and continuous path with values in $\mathbb{R}^d_+$. More precisely, we assume that the vector of capitalizations $X : [0, T] \to \mathbb{R}^d_+$, componentwise defined as the price $S_i(t)$ times the number of outstanding shares $n_i(t)$, is a deterministic and continuous path with values in $\mathbb{R}^d_+$, which exhibits pathwise continuous quadratic variation along $(\mathbb{T}_n)_{n\in\mathbb{N}}$ in the sense of [Föl81], i.e. we assume that $X : [0, T] \to \mathbb{R}^d_+$ satisfies Assumption 1.40.

Lemma 6.1. *The following are equivalent, where we abbreviate "admits continuous $\mathbb{S}_+^d$-valued quadratic variation along $(\mathbb{T}_n)_{n\in\mathbb{N}}$ in the sense of [Föl81]" by "admits quadratic variation".*

(i) $X : [0,T] \to \mathbb{R}_+^d$ *satisfies Assumption 1.40.*

(ii) $\log X$ *admits quadratic variation with* $[\log X_i, \log X_j]_t = \int_0^t \sigma_{ij}(s)ds$*, for all* $i,j = 1,...,d$ *and* $t \in [0,T]$*.*

(iii) $\mu : [0,T] \to \Delta^d$ *admits quadratic variation.*

(iv) $\log \mu$ *admits quadratic variation with* $[\log \mu_i, \log \mu_j]_t = \int_0^t \tau_{ij}(s)ds$*, for all* $i,j = 1,...,d$ *and* $t \in [0,T]$*, where*

$$\tau_{ij}(t) = \sigma_{ij}(t) - \sum_{k=1}^d \mu_k(t)\sigma_{kj}(t) - \sum_{l=1}^d \sigma_{il}(t)\mu_l(t) + \sum_{k,l=1}^d \mu_k(t)\sigma_{kl}(t)\mu_l(t),$$

for all $i,j = 1,...,d$ *and* $t \in [0,T]$*.*

Proof. First, we observe for some path $Z : [0,T] \to [0,\infty[$ of finite variation by [Son06, Proposition 2.2.10] that Z admits continuous $\mathbb{S}_+^d$-valued quadratic variation if and only if $\log Z$ does, and in this case, we have $[\log Z]_t = \int_0^t \frac{d[Z]_s}{Z(s)^2}$, for all $t \in [0,T]$. Now, suppose that (i) holds true, from which (ii) follows by using the polarization identity

$$[\log X_i, \log X_j]_t = \frac{1}{4}\left([\log X_i + \log X_j]_t - [\log X_i - \log X_j]_t\right)$$

and if we can show that $f(X_i(t), X_j(t))$ admits continuous $\mathbb{S}_+^d$-valued quadratic variation, where $f(x_i, x_j) = \log x_i + \log x_j$. Applying the pathwise Ito formula in Theorem 1.41 to $f(X_i(t), X_j(t))$ yields for all $t \in [0,T]$ and $i,j = 1,...,d$ that

$$\begin{aligned} f(X_i(t), X_j(t)) = f(X_i(0), X_j(0)) &+ \int_0^t \nabla f(X_i(s), X_j(s))^\top d \begin{pmatrix} X_i(s) \\ X_j(s) \end{pmatrix} \\ &+ \frac{1}{2}\sum_{k,l=1}^2 \int_0^t D_{kl} f(X_i(s), X_j(s)) d[X_{i_k}, X_{j_l}]_s, \end{aligned}$$

where $i_1 = j_1 = i$ and $i_2 = j_2 = j$. Hence, the quadratic variation of

$f(X_i(t), X_j(t))$ is well-defined because

$$\begin{aligned}[f(X_i(\cdot), X_j(\cdot))]_t &= \sum_{k,l=1}^{2} \int_0^t D_k f(X_i(s), X_j(s)) D_l f(X_i(s), X_j(s)) d[X_{i_k}, X_{i_l}]_s \\ &= \int_0^t \frac{d[X_i]_s}{X_i(s)^2} + \int_0^t \frac{d[X_j]_s}{X_j(s)^2} + 2\int_0^t \frac{d[X_i, X_j]_s}{X_i(s)X_j(s)},\end{aligned}$$

for all $i, j = 1, ..., d$ and $t \in [0, T]$, which implies (ii). The reverse direction (ii) $\Rightarrow$ (i) follows by a similar argument involving the exponential function establishing the equivalence (i) $\Leftrightarrow$ (ii). For the equivalence (ii) $\Leftrightarrow$ (iv), we observe by the polarization identity that $t \mapsto Z_\mu(t) = X_1(t) + ... + X_d(t)$ admits continuous $\mathbb{S}_+^d$-valued quadratic variation. Using the above argument, we conclude that $\log X_i$ and $\log Z_\mu$ both admit continuous $\mathbb{S}_+^d$-valued quadratic variation, from which it follows again by the polarization identity that $t \mapsto \log \mu_i(t) = \log X_i(t) - \log Z_\mu(t)$ admits continuous $\mathbb{S}_+^d$-valued quadratic variation. Moreover, we conclude by using the pathwise Ito formula in Theorem 1.41 that

$$\begin{aligned}[\log \mu_i, \log \mu_j]_t &= [\log X_i, \log X_j]_t - [\log Z_\mu, \log X_j]_t - [\log X_i, \log Z_\mu]_t \\ &\quad + [\log Z_\mu, \log Z_\mu]_t \\ &= \sigma_{ij}(t) - \sum_{k=1}^{d} \mu_k(t)\sigma_{kj}(t) - \sum_{l=1}^{d} \sigma_{il}(t)\mu_l(t) \\ &\quad + \sum_{k,l=1}^{d} \mu_k(t)\sigma_{kl}(t)\mu_l(t),\end{aligned}$$

for all $i, j = 1, ..., d$ and $t \in [0, T]$, which implies (iv). The reverse direction follows by similar arguments using the polarization identity for $t \mapsto \log X_i(t) = \log \mu_i(t) + \log Z_\mu(t)$. Finally the equivalence (iii) $\Leftrightarrow$ (iv) follows by similar arguments used for the equivalence (i) $\Leftrightarrow$ (ii). $\square$

In the following, we transfer the main result about functionally generated portfolios in [Fer02, Theorem 3.1.5] to this pathwise setting, which was obtained in [SSV18] for the path-dependent setting. In particular, we need to be careful when applying Theorem 1.41 for the Föllmer integral, where the integrand must be admissible such that the Föllmer integral exists. As in the previous chapter, we define the investment strategies as portfolios, which describe the fraction of wealth invested into each stock.

Definition 6.2. *A function $\pi : [0,T] \to \mathbb{R}^d$ is called a* portfolio *if it is an admissible integrand with respect to $\log X$ and if $\pi_1(t) + ... + \pi_d(t) = 1$, for all $t \in [0,T]$. In addition, a portfolio π is called* long-only *if $\pi_i(t) \geqslant 0$, for all $i = 1, ..., d$ and $t \in [0,T]$.*

Remark 6.3. *The market portfolio μ forms a long-only portfolio, because $\mu_1(t) + ... + \mu_d(t) = 1$ and $\mu_i(t) \geqslant 0$, for all $i = 1, ..., d$ and $t \in [0,T]$. Choosing the function $f(x) = \log(e^{x_1} + ... + e^{x_d})$, we see that $\nabla f(\log X(t)) = \mu(t)$, for all $t \in [0,T]$, such that μ is an admissible integrand on Δ^d for $\log X$.*

In order to define the wealth $Z_\pi : [0,T] \to \mathbb{R}$ of a portfolio $\pi : [0,T] \to \mathbb{R}^d$, we first observe that the number of shares $h : [0,T] \to \mathbb{R}^d$ invested in each stock is given by $h_i(t) = \frac{\pi_i(t) Z_\pi(t)}{X_i(t)}$, for $i = 1, ..., d$ and $t \in [0,T]$. Then, $h : [0,T] \to \mathbb{R}^d$ must be a self-financing trading strategy in the sense that changes in the portfolio wealth are only implied by trading in the market, which means that

$$Z_\pi(t) = Z_\pi(0) + \int_0^t h(s)^\top dX(s), \tag{6.1.1}$$

for all $t \in [0,T]$. In other words, the self-financing condition ensures that no external money is added to and no money is withdrawn from the portfolio wealth. Because of the self-financing condition in (6.1.1), we obtain the following expression for the wealth of a portfolio $\pi : [0,T] \to \mathbb{R}^d$.

Lemma 6.4. *For a portfolio $\pi : [0,T] \to \mathbb{R}^d$, the wealth process $Z_\pi : [0,T] \to \mathbb{R}$ satisfies*

$$\begin{aligned} \log Z_\pi(t) = \log Z_\pi(0) &+ \int_0^t \pi(s)^\top d \log X(s) + \frac{1}{2} \sum_{i=1}^d \int_0^t \pi_i(s) \sigma_{ii}(s) ds \\ &- \frac{1}{2} \sum_{i,j=1}^d \int_0^t \pi_i(s) \sigma_{ij}(s) \pi_j(s) ds, \end{aligned} \tag{6.1.2}$$

for all $t \in [0,T]$.

Proof. Since $\pi : [0,T] \to \mathbb{R}^d$ is a portfolio, it is an admissible integrand for $\log X$ such that the Föllmer integral in (6.1.2) exists. The other two integrals in (6.1.2) exist as Riemann-Stieltjes integrals, because $\pi : [0,T] \to \mathbb{R}^d$ is an admissible integrand for $\log X$ and in particular a continuous function of t.

Then, the logarithmic wealth process $\log Z_\pi : [0,T] \to \mathbb{R}$ in (6.1.2) is well-defined and it follows by combining the pathwise Ito formula in Theorem 1.41 and the self-financing condition in (6.1.1) that

$$\begin{aligned}\log Z_\pi(t) - \log Z_\pi(0) &= \int_0^t \frac{dZ_\pi(s)}{Z_\pi(s)} - \frac{1}{2}\int_0^t \frac{d[Z_\pi]_s}{Z_\pi(s)^2} \\ &= \int_0^t \left(\frac{\pi(s)}{X(s)}\right)^\top dX(s) - \frac{1}{2}\sum_{i,j=1}^d \int_0^t \frac{\pi_i(s)\pi_j(s)}{X_i(s)X_j(s)} d[X_i, X_j]_s,\end{aligned}$$

for all $t \in [0,T]$. In this case, the conclusion follows by another application of the pathwise Ito formula on $f(X(t)) = \log(X(t))$ as well as using the definition of $\sigma : [0,T] \to \mathbb{S}_+^d$ in Lemma 6.1(ii). □

With the notion of a portfolio $\pi : [0,T] \to \mathbb{R}^d$ and the portfolio wealth $Z_\pi : [0,T] \to \mathbb{R}$ in the pathwise setting, we introduce functionally generated portfolios as in the previous chapter.

6.2 Functionally Generated Portfolios

The main result of [Fer02] for functionally generated portfolios is the pathwise decomposition of the logarithmic relative return into a part involving the generating function and an additional drift process, which is captured by the so-called Fernholz' master equation. We say that a positive C^2-function $S : \Delta^d \to [0,\infty[$ *generates* the portfolio π if there exists a path of finite variation $\Theta : [0,T] \to \mathbb{R}$ such that

$$\log\left(\frac{Z_\pi(t)}{Z_\mu(t)}\right) = \log S(\mu(t)) + \Theta(t), \tag{6.2.1}$$

for all $t \in [0,T]$, where the generating function component dominates the short-term behaviour and the drift process determines the long-term performance. In the following, we adapt the main result [Fer02, Theorem 3.1.5] stated in Theorem 5.18 to this pathwise setting.

Theorem 6.5. *Let $S : U \to [0,\infty[$ be a twice differentiable function defined on an open neighborhood U of Δ^d such that $x \mapsto x_i D_i \log S(x)$ is bounded on Δ^d. Then $S : U \to [0,\infty[$ generates the portfolio $\pi : [0,T] \to \mathbb{R}^d$ with weights given by*

$$\pi_i(t) = \left(D_i \log S(\mu(t)) + 1 - \sum_{j=1}^{d} \mu_j(t) D_j \log S(\mu(t)) \right) \mu_i(t), \tag{6.2.2}$$

for all $i = 1, ..., d$ and $t \in [0,T]$, and the corresponding drift $\Theta : [0,T] \to \mathbb{R}$ is given by

$$\Theta(t) = -\sum_{i,j=1}^{d} \int_0^t \frac{1}{2S(\mu(s))} D_{ij} S(\mu(s)) \mu_i(s) \tau_{ij}(s) \mu_j(s) ds, \tag{6.2.3}$$

for all $t \in [0,T]$.

Proof. Let $\pi : [0,T] \to \mathbb{R}^d$ be the path with components defined in (6.2.2) such that $\sum_{i=1}^d \pi_i(t) = 1$, for all $t \in [0,T]$. Define $f(x) = \log S(\nabla g(x)) + g(x)$, with $g(x) = \log(e^{x_1} + ... + e^{x_d})$ such that $\nabla g(\log X(t)) = \mu(t)$, for all $t \in [0,T]$, for which it follows by the chain rule that

$$D_i f(x) = \sum_{j=1}^{d} D_j \log S(\nabla g(x)) D_{ji} g(x) + D_i g(x).$$

Hence, we have $\pi_i(t) = D_i f(\log X(t))$, for all $i = 1, ..., d$ and $t \in [0,T]$, which shows that $\pi : [0,T] \to \mathbb{R}^d$ is an admissible integrand for $\log X$ and therefore indeed a portfolio. Using Lemma 6.4 and subtracting $\log Z_\mu(t) = \log Z_\mu(0) + \int_0^t d \log Z_\mu(s)$ on both sides, it follows that

$$\begin{aligned} \log\left(\frac{Z_\pi(t)}{Z_\mu(t)}\right) = \log\left(\frac{Z_\pi(0)}{Z_\mu(0)}\right) + \sum_{i=1}^{d} \int_0^t \pi_i(s) d \log \mu_i(s) + \frac{1}{2} \sum_{i=1}^{d} \int_0^t \pi_i(s) \sigma_{ii}(s) ds \\ - \frac{1}{2} \sum_{i,j=1}^{d} \int_0^t \pi_i(s) \sigma_{ij}(s) \pi_j(s) ds, \end{aligned}$$

for all $t \in [0,T]$. Applying the pathwise Ito formula in Theorem 1.41 such that $d \log \mu_i(t) = d\mu_i(t)/\mu_i(t) - d[\mu_i]_t / 2\mu_i(t)^2$, for all $i = 1, ..., d$ and $t \in [0,T]$, using that the market weights sum up to one and $\mu : [0,T] \to \mathbb{R}^d$ lies by Lemma 5.6

in the null space of $\tau : [0,T] \to \mathbb{S}_+^d$, we conclude that

$$\log\left(\frac{Z_\pi(t)}{Z_\mu(t)}\right) = \log\left(\frac{Z_\pi(0)}{Z_\mu(0)}\right) + \sum_{i=1}^d \int_0^t D_i \log S(\mu(s)) d\mu_i(s) - \frac{1}{2}\sum_{i,j=1}^d \int_0^t D_i \log S(\mu(s)) D_j \log S(\mu(s)) \mu_i(s)\tau_{ij}(s)\mu_j(s) ds, \tag{6.2.4}$$

for all $t \in [0,T]$. On the other hand, applying the pathwise Ito formula in Theorem 1.41 to $\log S(\mu(t))$, which is possible because $\mu : [0,T] \to \mathbb{R}^d$ admits by Lemma 6.1 continuous $\mathbb{S}_+^d$-valued quadratic variation, and using that $D_{ij}\log S(x) = D_{ij}S(x)/S(x) - D_i \log S(x) D_j \log S(x)$, for $i,j = 1,...,d$, we observe that

$$\begin{aligned}\log S(\mu(t)) - \log S(\mu(0)) &= \sum_{i=1}^d \int_0^t D_i \log S(\mu(s)) d\mu_i(s) \\ &+ \frac{1}{2}\sum_{i,j=1}^d \int_0^t \frac{D_{ij}S(\mu(s))}{2S(\mu(s))} \mu_i(t)\tau_{ij}(t)\mu_j(t) dt \\ &- \frac{1}{2}\sum_{i,j=1}^d \int_0^t D_i \log S(\mu(s)) D_j \log S(\mu(s)) \mu_i(t)\tau_{ij}(t)\mu_j(t) dt,\end{aligned} \tag{6.2.5}$$

for all $t \in [0,T]$, where we have used Lemma 6.1 such that $d[\mu_i,\mu_j]_t = \mu_i(t)\tau_{ij}(t)\mu_j(t)dt$, for all $i,j = 1,...,d$ and $t \in [0,T]$. Finally, by subtracting (6.2.5) from (6.2.4), rearranging terms as well as using $\log(Z_\pi(0)/Z_\mu(0)) = \log S(\mu(0))$, we obtain the conclusion with drift $\Theta : [0,T] \to \mathbb{R}$ defined in (6.2.3). □

In addition, [Str12] extended Theorem 5.18 to portfolios generated by a function which can depend on an additional continuous path of finite variation $A : [0,T] \to \mathbb{R}^m$. Despite the explicit construction, it is not clear which economic process should be added for A and therefore the extended setting has so far been applied in [RX18], but in the probabilistic setting within the additive model introduced in [KR17].

Theorem 5.18 shows that a functionally generated portfolio π and its investment characteristics are fully determined by its generating function $S : U \to [0,\infty[$. As before, we index the portfolio and the drift process by its generating function and write π^S for the portfolio generated by S with drift Θ^S.

6.3 Path-Dependent Setting

The concept of functionally generated portfolios within the pathwise framework of SPT was extended in [SSV18] to the path-dependent setting such that portfolios can also depend on the previous values of the market weights, which is not the case in the usual pathwise setting of SPT described in Theorem 6.5. For this section, we follow along the lines of [SSV18, Section 2.2].

Definition 6.6. *We say that* $\pi : \Lambda_T \to \mathbb{R}^d$ *is a* functional portfolio *if it is an admissible functional integrand for* $\log X$ *on* $\mathbb{R}^d$ *such that* $\pi_1(t, \log X) + ... + \pi_d(t, \log X) = 1$*, for all* $t \in [0, T]$*, and if* $t \mapsto \int_0^t \pi(s, \log X) d \log X(s)$ *admits along* $(\mathbb{T}_n)_{n\in\mathbb{N}}$ *the quadratic variation*

$$\left[\int_0^{\cdot} \pi(s, \log X) d \log X(s)\right]_t = \sum_{i,j=1}^{d} \int_0^t \pi_i(s, \log X)\sigma_{ij}(s)\pi_j(s, \log X) ds, \quad (6.3.1)$$

for all $t \in [0, T]$.

Remark 6.7. *While the condition* (6.3.1) *is always satisfied in the case without path-dependence, it is in general not fulfilled in this path-dependent setting, where [AC17, Theorem 2.1] gives sufficient conditions for a functional integrand such that* (6.3.1) *is satisfied.*

Remark 6.8. *Choosing* $F(t, Y) = \log(e^{Y_1(t)} + ... + e^{Y_d(t)})$*, it follows that* $\nabla F(t, \log X) = \mu(t)$*, such that the market portfolio* μ *is a functional portfolio. Hereby,* (6.3.1) *holds true by [AC17, Theorem 2.1] and using that* $F \in \mathbb{C}^{1,2}(\Lambda_T)$ *is boundedness-preserving and uniformly Lipschitz continuous in* Y *with boundedness-preserving first vertical derivative* $\nabla F \in \mathbb{C}^{1,1}(\Lambda_T)$.

We now define the wealth of a functional portfolio $\pi : \Lambda_T \to \mathbb{R}^d$. Inspired by the setting without path-dependence described in Lemma 6.4, we define the

wealth of π as a path $Z_\pi : [0,T] \to \mathbb{R}$ given by

$$\begin{aligned}\log Z_\pi(t) = \log Z_\pi(0) &+ \sum_{i=1}^{d} \int_0^t \pi_i(s, \log X) d \log X_i(s) \\ &+ \frac{1}{2} \sum_{i=1}^{d} \int_0^t \pi_i(s, \log X) \sigma_{ii}(s) ds \\ &- \frac{1}{2} \sum_{i,j=1}^{d} \int_0^t \pi_i(s, \log X) \pi_j(s, \log X) \sigma_{ij}(s) ds,\end{aligned}$$

for all $t \in [0,T]$, where $Z_\pi(0)$ denotes the initial endowment of π. In addition, we say that a non-anticipative functional F *generates* the functional portfolio $\pi : \Lambda_T \to \mathbb{R}^d$ if there exists a path of finite variation $\Theta : [0,T] \to \mathbb{R}$ such that

$$\log \left(\frac{Z_\pi(t)}{Z_\mu(t)} \right) = \log F(t, \mu) + \Theta(t),$$

for all $t \in [0,T]$. We recall the main result of [SSV18], which describes the weights of a portfolio generated by the vertical derivatives of the smooth non-anticipative functional $F \in \mathbb{C}^{1,2}(\Lambda_T)$. In the same spirit as before, we index sometimes the generated portfolio $\pi^F : \Lambda_T \to \mathbb{R}^d$ by its generating functional.

Theorem 6.9 ([SSV18, Theorem 2.16])**.** *Let $F \in \mathbb{C}^{1,2}(\Lambda_T)$ be a strictly positive non-anticipative functional, which is boundedness-preserving and locally path-continuous. Then, $F : \Lambda_T \to \mathbb{R}$ generates the functional portfolio $\pi : \Lambda_T \to \mathbb{R}^d$ with weights given by*

$$\pi_i(t, \mu) = \left(\partial_i \log F(t, \mu) + 1 - \sum_{j=1}^{d} \mu_j(t) \partial_j \log F(t, \mu) \right) \mu_i(t),$$

for all $i = 1, ..., d$ and $t \in [0,T]$, such that

$$\Theta(t) = - \sum_{i,j=1}^{d} \int_0^t \frac{1}{2F(s, \mu)} \partial_{ij} F(s, \mu) \mu_i(s) \mu_j(s) \tau_{ij}(s) ds - \int_0^t \frac{\mathcal{D}F(s, \mu)}{F(s, \mu)} ds,$$

for all $t \in [0,T]$.

Altogether, the concept of functionally generated portfolios can be used in the probabilistic setting (Chapter 5) or in the pathwise setting (this chapter), in both of which the extension to the path-dependent setting is possible. The advantage of the pathwise setting is that it does not depend on a partic-

ular choice of the stochastic process, which models the capitalization vector. Moreover, one can usually only observe one historical trajectory of a real equity market, which corresponds to one state of the world, where functionally generated portfolios can still be described within this pathwise setting.

Chapter 7

Learning an Optimal Portfolio

As we have seen in the previous chapter, investment strategies can be constructed in the pathwise framework of Stochastic Portfolio Theory (SPT) by the choice of a generating function. In this chapter, we will show that conventional neural networks between finite dimensional spaces and functional neural networks on the space of stopped paths can be used to learn a functionally generated portfolio in both, the classical setting of SPT and in the path-dependent setting. In order to learn a portfolio, we define the objective of an investor as a performance functional, which measures the "goodness" of the portfolio according to the objective of the investor.

For the whole chapter, we assume the pathwise setting of SPT such that paths of the vector of capitalizations $X : [0,T] \to \mathbb{R}^d_+$ are continuous with non-negative values, where $d \in \mathbb{N}$ denotes the number of assets. In order to apply the pathwise setting of functionally generated portfolios, we assume throughout this chapter that $X : [0,T] \to \mathbb{R}^d_+$ exhibits pathwise continuous quadratic variation along a sequence of partitions $(\mathbb{T}_n)_{n\in\mathbb{N}}$ in the sense of [Föl81], i.e. that $X : [0,T] \to \mathbb{R}^d_+$ satisfies Assumption 1.40.

7.1 Approximation of Functionally Generated Portfolio

In this section, we show that every functionally generated portfolio can be approximated by a portfolio, which is generated by a neural network $\Phi : \Delta^d \to \mathbb{R}$, where the neural network approximates the generating function of the optimal portfolio. More precisely, we approximate the strictly positive generating function $S : \Delta^d \to [0,\infty[$ by the exponential of a neural network $\Phi : \mathbb{R}^d \to \mathbb{R}$ to account for the positivity of the generating function, i.e. $\Phi : \mathbb{R}^d \to \mathbb{R}$ should approximate $\log S : \Delta^d \to \mathbb{R}$.

For the rest of this chapter, we fix a twice differentiable function $\rho : \mathbb{R} \to \mathbb{R}$ such that $\mathcal{NN}^{\rho}_{d,\infty}$ is dense in $C^2(\mathbb{R}^n)$ on compacta, and denote by $(\mathcal{NN}_M)_{M\in\mathbb{N}}$ a sequence of subsets of $\mathcal{NN}^{\rho}_{d,\infty}$ such that

$$\mathcal{NN}_M \subseteq \mathcal{NN}_{M+1}, \quad \bigcup_{M\in\mathbb{N}} \mathcal{NN}_M = \mathcal{NN}^{\rho}_{d,\infty}, \quad \mathcal{NN}_M = \left\{\Phi^\theta;\, \theta \in \Theta_{d,M}\right\}, \tag{7.1.1}$$

for all $M \in \mathbb{N}$, where $\Theta_{d,M} \subseteq \mathbb{R}^q$, for some $q \in \mathbb{N}$ (depending on M). We have two particular examples in mind: we could define $\mathcal{NN}_M$ as the neural networks in $\mathcal{NN}^{\rho}_{d,\infty}$ with arbitrary number of layers and nodes, but with connectivity at most M. Otherwise, we could take for $\mathcal{NN}_M$ the set of all neural networks with "fixed architecture", i.e. with non-decreasing sequences $(L^{(M)})_{M\in\mathbb{N}}$ and $(N_1^{(M)})_{M\in\mathbb{N}}, ..., (N_{L^{(M)}}^{(M)})_{M\in\mathbb{N}}$.

Moreover, by using the equivalence of the 1-norm and 2-norm, and the parallelogram inequality, we observe that (BV_κ) implies

$$x^\top \tau(t) x \leqslant 2\kappa(d+1)\|x\|^2, \tag{7.1.2}$$

for all $x \in \mathbb{R}^d$ and $t \in [0,T]$.

Using the sequence $(\mathcal{NN}_M)_{M\in\mathbb{N}}$ of neural network subsets satisfying (7.1.1) and the universal approximation result with derivatives in Theorem 2.6, we can show that every functionally generated portfolio can be approximated arbitrarily well by the portfolio generated from a sufficiently large neural network.

Theorem 7.1. *Let $S : \Delta^d \to [0,\infty[$ with $\log S \in C^2(\overline{\Delta}^d)$ generate the portfolio π^S. Assume (BV_κ) and suppose that $\rho \in C^2(\mathbb{R})$ is such that $\mathcal{NN}^{\rho}_1$ is dense in $C^2(K)$, for all compact subsets $K \subset \mathbb{R}$. Then, for every $\varepsilon > 0$, there exists some $M \in \mathbb{N}$ and $\Phi \in \mathcal{NN}_M$ generating the portfolio $\pi^{\exp(\Phi)}$ such that*

$$\left|\pi_i^S(t) - \pi_i^{\exp(\Phi)}(t)\right| < \varepsilon, \tag{7.1.3}$$

for all $i = 1, ..., d$ and $t \in [0,T]$. Furthermore, it follows that

$$\left|\log\left(Z_{\pi^S}(t)\right) - \log\left(Z_{\pi^{\exp(\Phi)}}(t)\right)\right| < (1+Ct)\,\varepsilon, \tag{7.1.4}$$

for all $t \in [0,T]$ and some constant $C > 0$.

Proof. Since $\log S \in C^2(\overline{\Delta}^d)$, it follows from Theorem 3.52 that $\log S$ can be approximated by a neural network in $C^2(\overline{\Delta}^d)$. Hence, for some fixed $\varepsilon > 0$

there exists a number $M \in \mathbb{N}$ such that

$$\begin{aligned} |\log S(x) - \Phi(x)| &< \frac{\varepsilon}{2}, \\ |D_i \log S(x) - D_i\Phi(x)| &< \frac{\varepsilon}{2}, \\ |D_{ij} \log S(x) - D_{ij}\Phi(x)| &< \frac{\varepsilon}{2}, \end{aligned} \tag{7.1.5}$$

for all $i, j = 1, ..., d$ and $x \in \overline{\Delta}^d$, where $\Phi \in \mathcal{NN}^{\rho}_{M,d}$. For (7.1.3), we use Theorem 5.18, the triangle inequality and (7.1.5) to conclude that

$$\begin{aligned} \left|\pi_i^S(t) - \pi_i^{\exp(\Phi)}(t)\right| &\leqslant |D_i \log S(\mu(t)) - D_i\Phi(\mu(t))|\, \mu_i(t) \\ &\quad + \sum_{j=1}^{d} \mu_j(t)\, |D_j \log S(\mu(t)) - D_j\Phi(\mu(t))|\, \mu_i(t) < \varepsilon, \end{aligned}$$

for all $i = 1, ..., d$ and $t \in [0, T]$, where we have used that $0 \leqslant \mu_i(t) \leqslant 1$ and $\sum_{i=1}^d \mu_i(t) = 1$. For (7.1.4), we use (6.2.1) and insert (7.1.5) such that

$$\begin{aligned} |\log(Z_{\pi^S}(t)) - \log(Z_{\pi^{\exp(\Phi)}}(t))| &= \left|\log\left(\frac{Z_{\pi^S}(t)}{Z_\mu(t)}\right) - \log\left(\frac{Z_{\pi^{\exp(\Phi)}}(t)}{Z_\mu(t)}\right)\right| \\ &\leqslant |\log S(\mu(t)) - \Phi(\mu(t))| + \left|\Theta^S(t) - \Theta^{\exp(\Phi)}(t)\right| \\ &< \frac{\varepsilon}{2} + \left|\Theta^S(t) - \Theta^{\exp(\Phi)}(t)\right|, \end{aligned} \tag{7.1.6}$$

for all $t \in [0, T]$. Using the definition of the drift process in Theorem 6.5 and $D_{ij}S(x)/S(x) = D_{ij} \log S(x) + D_i \log S(x) D_j \log S(x)$, it follows by applying the triangle inequality and (BV_κ) implying (7.1.2) that

$$\begin{aligned} \left|\Theta^S(t) - \Theta^{\exp(\Phi)}(t)\right| &\leqslant \sup_{\substack{x \in \overline{\Delta}^d \\ i,j=1,\dots,d}} |D_{ij} \log S(x) - D_{ij}\Phi(x)|\, \kappa(d+1)t \\ &\quad + \sup_{\substack{x \in \overline{\Delta}^d \\ i,j=1,\dots,d}} |D_i \log S(x) D_j \log S(x) - D_i\Phi(x) D_j\Phi(x)|\, \kappa(d+1)t \\ &< (1 + C_1 + C_2)\kappa(d+1)t\frac{\varepsilon}{2}, \end{aligned} \tag{7.1.7}$$

for all $t \in [0, T]$, where we have used $a_i a_j - b_i b_j = a_i(a_j - b_j) + b_j(a_i - b_i)$ with $a_i = D_i \log S(x)$ and $b_i = D_i\Phi(x)$, and where $C_1 = \sup_{x \in \overline{\Delta}^d,\, i=1,\dots,d} |D_i \log S(x)|$ and $C_2 = \sup_{x \in \overline{\Delta}^d,\, j=1,\dots,d} |D_j\Phi(x)|$ are finite. Finally, inserting (7.1.7) into (7.1.6) and rescaling $\varepsilon > 0$ yields the conclusion. □

Theorem 7.1 indicates that a functionally generated portfolio can be approximated arbitrarily well by a portfolio generated from a sufficiently large neural network, where the difference of the logarithmic wealth increases linearly in time due to the approximation error of the portfolio weights. However, having finite terminal time, the neural network can be enlarged to get a better approximation of the terminal wealth.

In the path-dependent setting of SPT, where portfolios are generated by smooth non-anticipative functionals, we use functional neural networks introduced in Chapter 4 to approximate the non-anticipative functional of the optimal portfolio. For this purpose, we consider the space of stopped α-Hölder continuous paths $(\Lambda_T^\alpha, d_\infty)$ and define the admissible weight function $\psi : \Lambda_T^\alpha \to]0, \infty[$ as in Remark 4.5 such that the stopped path of $X : [0,T] \to \mathbb{R}_+^d$ belongs to a compact subset of Λ_T^α. Using the weighted space (Λ_T^α, ψ), we define the space of weighted smooth functionals $\mathcal{L}\mathrm{ip}_\psi^{1,2}(\Lambda_T^\alpha)$ as the $\|\cdot\|_{\psi,1,2}$-completion of $\mathcal{L}\mathrm{ip}_b^{1,2}(\Lambda_T^\alpha)$, where the weighted norm $\|\cdot\|_{\psi,1,2}$ is given by

$$\begin{aligned}\|F\|_{\psi,1,2} = &\sup_{(t,Y)\in\Lambda_T^\alpha} \frac{|F(t,Y)|}{\psi(t,Y)} + \sup_{(t,Y)\in\Lambda_T^\alpha} \frac{|\mathcal{D}F(t,Y)|}{\psi(t,Y)} \\ &+ \sum_{1\leqslant|\mathbf{m}|\leqslant 2} \sup_{(t,Y)\in\Lambda_T^\alpha} \frac{|\partial_{\mathbf{m}}F(t,Y)|}{\psi(t,Y)},\end{aligned}$$

for $F \in \mathcal{L}\mathrm{ip}_\psi^{1,2}(\Lambda_T^\alpha)$. Then, the universal approximation theorem for smooth non-anticipative functionals in Theorem 4.6 shows that every path-dependent functionally generated portfolio can be approximated by a portfolio, which is generated by a functional neural network, where we assume that $\{\mathcal{FN}_M\}_{M\in\mathbb{N}}$ is an increasing sequence of subsets in $\mathcal{FN}_{d,\infty}^\rho$ such that $\bigcup_{M\in\mathbb{N}} \mathcal{FN}_M = \mathcal{FN}_{d,\infty}^\rho$.

Theorem 7.2. *Let the strictly positive $F \in \mathbb{C}^{1,2}(\Lambda_T^\alpha)$ generate the portfolio π^F such that $\log F \in \mathcal{L}ip_\psi^{1,2}(\Lambda_T^\alpha)$. Assume (BV_κ) and suppose that $\rho \in \mathcal{L}ip_b^2(\mathbb{R})$ is such that $\mathcal{NN}_1^\rho$ is dense in $C^2(K)$, for all compact subsets $K \subset \mathbb{R}$. Then, for every $\varepsilon > 0$ there exists some $M \in \mathbb{N}$ and a functional neural network $\Phi \in \mathcal{FN}_M$ generating the portfolio $\pi^{\exp(\Phi)}$ such that*

$$\left|\pi_i^F(t) - \pi_i^{\exp(\Phi)}(t)\right| < \varepsilon,$$

for all $i = 1, ..., d$ and $t \in [0,T]$, and

$$|\log(Z_{\pi^F}(t)) - \log(Z_{\pi^{\exp(\Phi)}})| < (1 + Ct)\varepsilon,$$

for all $t \in [0,T]$ and some constant $C > 0$.

Proof. Since $X : [0,T] \to \mathbb{R}^d_+$ satisfies Assumption 1.40 there exists by Remark 4.5 some sufficiently large $R > 0$ such that $\psi(t,\mu) \leqslant R$, for all $t \in [0,T]$. As $\log F \in \mathcal{L}\mathrm{ip}^{1,2}_{\psi}(\Lambda^{\alpha}_T)$, we can apply Theorem 4.6 to conclude that there exists for fixed $\varepsilon > 0$ some $\Phi \in \mathcal{FN}^{\rho}_{d,M}$ satisfying $\|\log F - \Phi\|_{\psi,p,q} \leqslant \frac{\varepsilon}{2R}$. From this, it follows that

$$|\log F(t,\mu) - \Phi(t,\mu)| \leqslant \psi(t,\mu)\|\log F - \Phi\|_{\psi,p,q} < \frac{\varepsilon}{2},$$

for all $t \in [0,T]$, and similarly $|\mathcal{D}\log F(t,\mu) - \mathcal{D}\Phi(t,\mu)| < \frac{\varepsilon}{2}$ as well as $|\partial_{ij}\log F(t,\mu) - \partial_{ij}\Phi(t,\mu)| < \frac{\varepsilon}{2}$, for all $1 \leqslant |\mathbf{m}| \leqslant 2$ and $t \in [0,T]$. Then, the expression for the portfolio weights in Theorem 6.9 implies that

$$\begin{aligned}\left|\pi^F_i(t) - \pi^{\exp(\Phi)}_i(t)\right| &\leqslant |\partial_i \log F(t,\mu) - \partial_i\Phi(t,\mu)|\,\mu_i(t) \\ &\quad + \sum_{j=1}^d \mu_j(t)\,|\partial_j \log F(t,\mu) - \partial_j\Phi(t,\mu)|\,\mu_i(t) < \varepsilon,\end{aligned}$$

for all $i = 1,...,d$ and $t \in [0,T]$. The approximation of the wealth follows then by the same arguments as in the proof of Corollary 7.1, where we use the drift in Theorem 6.9 for the approximation of the logarithmic relative return and the chain rule in [SV18, Lemma 3.3] such that $\mathcal{D}F(s,\mu)/F(s,\mu) = \mathcal{D}\log F(s,\mu)$, for all $t \in [0,T]$. □

In both cases, the classical and path-dependent setting of SPT, we obtain a parametric family of functionally generated portfolios obtained from (functional) neural networks, which can be used to learn the optimal portfolio by learning the optimal weights in the corresponding (functional) neural network.

7.2 Performance Functional

In order to train "good" investment strategies, we need to define the investor's objective, which is encoded in a user-specified preformance functional and was first introduced within the framework of SPT in [KSV16]. In addition, we introduce the set of functionally generated portfolios, from which the optimal portfolio should be chosen, given by

$$\begin{aligned}\mathcal{FG} = \Bigg\{ &\pi^S : \Delta^d \to \mathbb{R}^d \text{ defined by} \\ &\pi_i(x) = \left(D_i \log S(x) + 1 - \sum_{j=1}^d x_j D_j \log S(x)\right) x_i;\ S \in C^2(\Delta^d)\Bigg\},\end{aligned}$$

where a generic element $\pi \in \mathcal{FG}$ denotes by a slight abuse of notation a *portfolio map*, yielding a portfolio with $\pi(t) = \pi(\mu(t))$, for $t \in [0,T]$. We equip the space $\mathcal{FG}$ with the metric $d_{\mathcal{FG}}(\pi,\eta) = \max_{i=1,\dots,d} \sup_{x\in\Delta^d} |\pi_i(x) - \eta_i(x)|$, for $\pi,\eta \in \mathcal{FG}$, and observe that $\mathcal{FG}$ is convex, but neither totally bounded nor separable, see [Won15].

Definition 7.3. *The map $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ is called a* performance functional *if it is lower semi-continuous, i.e. for every sequence $(\pi_n)_{n\in\mathbb{N}}$ in $\mathcal{FG}$ converging to $\pi \in \mathcal{FG}$, it holds that $\liminf_{n\to\infty} \mathcal{P}(\pi_n) \geqslant \mathcal{P}(\pi)$.*

Given a performance functional $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$, we can measure the performance of a portfolio π with respect to the given objective of the investor. If the objective is not reached, i.e. not minimized in the right way, the portfolio π needs to be adapted such that the value of $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ further decreases. Furthermore, we introduce the subset of bounded α-Hölder continuous generating functions given by

$$\mathcal{G}^{\alpha,b} = \left\{ S \in C^2(\overline{\Delta}^d);\ \|S\|_{C^{2,\alpha}} \leqslant b \text{ and } S \geqslant \frac{1}{b} \right\},$$

for some $\alpha, b > 0$, where the norm $\|\cdot\|_{C^{2,\alpha}}$ is defined as

$$\|S\|_{C^{2,\alpha}} = \max_{0\leqslant|\mathbf{m}|\leqslant 2} \sup_{x\in\overline{\Delta}^d} |D_{\mathbf{m}}S(x)| + \max_{0\leqslant|\mathbf{m}|\leqslant 2} \sup_{\substack{x,y\in\overline{\Delta}^d \\ x\neq y}} \frac{|D_{\mathbf{m}}S(x) - D_{\mathbf{m}}S(y)|}{\|x-y\|^\alpha}, \tag{7.2.1}$$

and where we omit the second term in (7.2.1) if $\alpha = 0$. The corresponding set of functionally generated portfolios is given by

$$\mathcal{FG}^{\alpha,b} = \left\{ \pi^S : \Delta^d \to \mathbb{R}^d;\ S \in \mathcal{G}^{\alpha,b} \right\}.$$

Using that $\mathcal{G}^{\alpha,b}$ is compact with respect to $\|\cdot\|_{C^{2,0}}$ and $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ is lower semi-continuous, we obtain the existence of at least one minimizer on the set of bounded Hölder continuous portfolio maps $\mathcal{FG}^{\alpha,b}$.

Lemma 7.4. *For every $\alpha, b > 0$, the set $\mathcal{FG}^{\alpha,b}$ is compact with respect to $d_{\mathcal{FG}}$. Hence, a performance functional $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ attains in particular its minimum on $\mathcal{FG}^{\alpha,b}$.*

Proof. Fix $\alpha, b > 0$ and observe that the map $S \mapsto \pi^S$ is continuous from $(\mathcal{G}^{\alpha,b}, \|\cdot\|_{C^{2,0}})$ to $(\mathcal{FG}^{\alpha,b}, d_{\mathcal{FG}})$. This can be seen by using the definition of a

functionally generated portfolio in Theorem 5.18 such that

$$\begin{aligned} d_{\mathcal{FG}}\left(\pi^{S_1}, \pi^{S_2}\right) &\leqslant 2 \max_{i=1,\dots,d} \sup_{x \in \Delta^d} |D_i \log S_1(x) - D_i \log S_2(x)| \\ &\leqslant 2b \max_{i=1,\dots,d} \sup_{x \in \overline{\Delta}^d} |D_i S_1(x) - D_i S_2(x)| \\ &\leqslant 2b \|S_1 - S_2\|_{C^{2,0}}, \end{aligned}$$

for all $S_1, S_2 \in \mathcal{G}^{\alpha,b}$, where we have used that the logarithm is Lipschitz continuous on $[1/b, \infty[$ with Lipschitz constant b. Since $\mathcal{G}^{\alpha,b}$ is by [CSW19, Lemma 4.1] compact with respect to $\|\cdot\|_{C^{2,0}}$ and $S \mapsto \pi^S$ is continuous, the image $\mathcal{FG}^{\alpha,b}$ is compact with respect to $d_{\mathcal{FG}}$. The existence of a minimizer follows then by [Kos10, Satz 3.1.15], where the lower semi-continuity of $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ is used. □

Lemma 7.4 ensures the existence of a minimizer at least among the restricted set $\mathcal{FG}^{\alpha,b}$ of bounded and Hölder continuous portfolio maps. For bigger spaces such as $\mathcal{FG}$, it is not clear if such a minimizer exists. However using that $\mathcal{FG}^{\alpha,b} \subset \mathcal{FG}^{\alpha',b'} \subset \mathcal{FG}$, for $\alpha' < \alpha$ and $b < b'$, we obtain a sequence $(\hat{\pi}_n)_{n\in\mathbb{N}}$ of minimizing portfolio maps $\hat{\pi}_n = \arg\min_{\pi \in \mathcal{FG}^{1/n,n}} \mathcal{P}(\pi)$, for $n \in \mathbb{N}$, whose performances decreases and converges to the infimum of $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ on $\mathcal{FG}$, if $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ is for example bounded from below.

In the following, we consider examples of performance functionals for different investment objectives, for example to learn an arbitrage strategy. We know that such arbitrage opportunities exist in sufficiently diverse and volatile markets, where for example the diversity-weighted portfolio of Example 5.19(v) can be used to outperform the market portfolio, see [FKK05]. Hereby, a portfolio π is called a *relative arbitrage* with respect to the benchmark portfolio η if there exists some $t^* \in [0, T]$ such that

$$\mathbb{P}[Z_\pi(t^*) \geqslant Z_\eta(t^*)] = 1 \qquad \text{and} \qquad \mathbb{P}[Z_\pi(t^*) > Z_\eta(t^*)] > 0.$$

The relative arbitrage π is called *strong* if $\mathbb{P}[Z_\pi(t^*) > Z_\eta(t^*)] = 1$. Combining this with Theorem 7.1, we therefore only need to train a neural network $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that the generated portfolio $\pi^{\exp(\Phi)}$ will be most likely an arbitrage strategy. On the other hand, we consider also other performance functionals, representing for example the hedging problem.

Proposition 7.5. *The following maps are performance functionals:*

(i) $\mathcal{P}^H(\pi) = \int_0^T l\left(Z_\pi(t) - z_t\right) dt$, *for some other path* $z : [0,T] \to \mathbb{R}$ *and for some lower semi-continuous function* $l : \mathbb{R} \to \mathbb{R}$.

(ii) $\mathcal{P}^R(\pi) = -\left(\log\left(\frac{Z_\pi(T)}{Z_\eta(T)}\right) - \log\left(\frac{Z_\pi(0)}{Z_\eta(0)}\right)\right)$, *for some other portfolio* $\eta : [0,T] \to \mathbb{R}^d$.

(iii) $\mathcal{P}^V(\pi) = \int_0^T \left(\pi(t) - \eta(t)\right)^\top \sigma(t) \left(\pi(t) - \eta(t)\right) dt$, *for some other portfolio* $\eta : [0,T] \to \mathbb{R}^d$.

Proof. In order to prove that $\mathcal{P}^R : \mathcal{FG} \to \mathbb{R}$ is a performance functional, we show that $\pi \mapsto \log\left(Z_\pi(t)\right)$ is continuous from $(\mathcal{FG}, d_{\mathcal{FG}})$ to $(\mathbb{R}, |\cdot|)$, for every $t \in [0,T]$. For this purpose, let $\pi_1, \pi_2 \in \mathcal{FG}$. Then, it follows from Lemma 6.4 and the stability estimate for rough integration in [FH14, Theorem 4.16] that

$$\begin{aligned}
\left|\mathcal{P}^R(\pi_1) - \mathcal{P}^R(\pi_2)\right| &\leqslant \left| \int_0^T \left(\pi_1(\mu(s)) - \pi_2(\mu(s))\right)^\top d \log X(s) \right| \\
&+ \frac{1}{2} \left| \int_0^T \left(\pi_1(\mu(s)) - \pi_2(\mu(s)\right)^\top \mathrm{diag}(\sigma(s)) ds \right| \\
&+ \frac{1}{2} \left| \int_0^T \left(\pi_1(\mu(s)) - \pi_2(\mu(s))\right)^\top \sigma(s) (\pi_1(\mu(s)) - \pi_2(\mu(s))) ds \right| \\
&\leqslant C d_{\mathcal{FG}}(\pi_1, \pi_2) + \frac{1}{2} d_{\mathcal{FG}}(\pi_1, \pi_2) \max_{i=1,\dots,d} [\log X_i, \log X_i]_T \\
&+ \frac{1}{2} d_{\mathcal{FG}}(\pi_1, \pi_2)^2 \max_{i,j=1,\dots,d} [\log X_i, \log X_j]_T,
\end{aligned}$$

for all $t \in [0,T]$ and some constant $C > 0$, which depends on the pathwise quadratic variation $[\log X_i, \log X_j]$, for $i, j = 1, ..., d$. From this, it follows that $\mathcal{P}^R : \mathcal{FG} \to \mathbb{R}$ is continuous, therefore also lower semi-continuous and a performance functional.

For $\mathcal{P}^H : \mathcal{FG} \to \mathbb{R}$, we use lower semi-continuity of l such that $\pi \mapsto l\left(e^{\log(Z_\pi(t)/Z_\mu(t)) + \log Z_\mu(t)} - z_t\right)$ is also lower semi-continuous. Hence, it follows by Fatou's lemma for $(\pi^n)_{n\in\mathbb{N}}$ converging to π that

$$\liminf_{n\to\infty} \mathcal{P}^H(\pi^n) = \liminf_{n\to\infty} \int_0^T l\left(Z_{\pi^n}(t) - z_t\right) dt \geqslant \int_0^T l\left(Z_\pi(t) - z_t\right) dt = \mathcal{P}^H(\pi),$$

such that $\mathcal{P}^H : \mathcal{FG} \to \mathbb{R}$ is a performance functional. For $\mathcal{P}^V : \mathcal{FG} \to \mathbb{R}$, we observe that the quadratic form $x \mapsto x^\top A x$ is continuous, for any matrix

$A \in \mathbb{R}^{d\times d}$, such that Fatou's lemma implies for $(\pi^n)_{n\in\mathbb{N}}$ converging to π that

$$\begin{aligned}\liminf_{n\to\infty} \mathcal{P}^V(\pi^n) &= \liminf_{n\to\infty} \int_0^T (\pi^n(t)-\eta(t))^\top \sigma(t)\,(\pi^n(t)-\eta(t))\,dt \\ &\geqslant \int_0^T (\pi(t)-\eta(t))^\top \sigma(t)\,(\pi(t)-\eta(t))\,dt \\ &= \mathcal{P}^V(\pi),\end{aligned}$$

which shows that $\mathcal{P}^V : \mathcal{FG} \to \mathbb{R}$ is as well a performance functional. □

For applications, $\mathcal{P}^H$ can be used for hedging a given value process $z = (z_t)_{t\in[0,T]}$ and recover therewith the generating function of the hedging strategy, whereas $\mathcal{P}^R$ and $\mathcal{P}^V$ compare the performance of the portfolio against the benchmark portfolio $\eta : [0,T] \to \mathbb{R}^d$, e.g. the market portfolio. In practice, $\mathcal{P}^R$ and $\mathcal{P}^V$ are usually combined in order to account for the return and the risk simultaneously, where we see that linear combinations of performance functionals are again performance functionals.

After having established the notion of a performance functional, we analyse if learning the parameters in the neural network indeed approximates the infimum of the given performance functional over all functionally generated portfolios. For this purpose, we use Theorem 7.2 such that every functionally generated portfolio can be approximated by a portfolio generated by a neural network.

Theorem 7.6. *Let $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ be a performance functional and assume that (BV_κ) holds true. Then*

$$\inf_{\pi\in\mathcal{FG}} \mathcal{P}(\pi) = \lim_{M\to\infty} \inf_{\Phi\in\mathcal{NN}_M} \mathcal{P}(\pi^{\exp(\Phi)}).$$

Proof. First, we observe that for every $M \in \mathbb{N}$ it holds that $\mathcal{NN}_M \subset \mathcal{NN}_{M+1} \subset \{\log S : \Delta^d \to \mathbb{R};\ S \in C^2(\Delta^d)\}$, which implies

$$\inf_{\pi\in\mathcal{FG}} \mathcal{P}(\pi) \leqslant \inf_{\Phi\in\mathcal{NN}_{M+1}} \mathcal{P}(\pi^{\exp(\Phi)}) \leqslant \inf_{\Phi\in\mathcal{NN}_M} \mathcal{P}(\pi^{\exp(\Phi)}),$$

for all $M \in \mathbb{N}$. Thus, it suffices to show that for every $\varepsilon > 0$ there exists some $M \in \mathbb{N}$ such that

$$\inf_{\pi\in\mathcal{FG}} \mathcal{P}(\pi) > \inf_{\Phi\in\mathcal{NN}_M} \mathcal{P}(\pi^{\exp(\Phi)}) - \varepsilon. \tag{7.2.2}$$

Fix some $\varepsilon > 0$ such that there exists by definition some $R : \Delta^d \to [0,\infty[$ with

$\log R \in C^2(\overline{\Delta}^d)$ satisfying

$$\inf_{\pi \in \mathcal{FG}} \mathcal{P}(\pi) \geqslant \mathcal{P}(\pi^R) - \frac{\varepsilon}{2}. \tag{7.2.3}$$

On the other hand, we find by Theorem 7.1 some neural network $\Phi \in \mathcal{NN}_M$ with $|\pi_i^R(t) - \pi_i^{\exp(\Phi)}(t)| < \varepsilon/2$, for all $i = 1, ..., d$ and $t \in [0, T]$. Using (7.2.3) and applying lower semi-continuity of $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ in the second step, it follows that

$$\inf_{\pi \in \mathcal{FG}} \mathcal{P}(\pi) \geqslant \mathcal{P}(\pi^R) - \frac{\varepsilon}{2} > \mathcal{P}(\pi^{\exp(\Phi)}) - \varepsilon \geqslant \inf_{\Phi \in \mathcal{NN}_M} \mathcal{P}(\pi^{\exp(\Phi)}) - \varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrarily, we obtain (7.2.2) and therefore the conclusion. □

Theorem 7.6 shows that the universal approximation of neural networks in Theorem 2.5 can be used to approximate an optimal functionally generated portfolios with respect to a given performance functional $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$, by increasing the connectivity of the neural network.

7.3 Path-Dependent Setting

In the same spirit as before, we want to learn "good investment maps" in the path-dependent setting of SPT, where we need to redefine the performance functional as the objective of the investor. We assume again the pathwise setting of SPT, where the vector of capitalizations $X : [0, T] \to \mathbb{R}^d_+$ satisfies Assumption 1.40. Then, we introduce the set of path-dependent functionally generated portfolios given by

$$\mathcal{FG} = \left\{ \pi^F ; \log F \in \mathcal{L}\mathrm{ip}^{1,2}_{\psi}(\Lambda^{\alpha}_T) \right\},$$

and equip it with the metric $d_{\mathcal{FG}}(\pi, \eta) = \max_{i=1,...,d} \sup_{(t,Y) \in \Lambda^{\alpha}_T} |\pi_i(t, Y) - \eta_i(t, Y)|$, for $\pi, \eta \in \mathcal{FG}$.

Definition 7.7. *The map $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ is a* performance functional *if it is lower semi-continuous, i.e. for every sequence $(\pi_n)_{n \in \mathbb{N}}$ in $\mathcal{FG}$ converging to $\pi \in \mathcal{FG}$, it holds that $\liminf_{n \to \infty} \mathcal{P}(\pi_n) \geqslant \mathcal{P}(\pi)$.*

By construction of the weighted space $(\Lambda^{\alpha}_T, \psi)$, a performance functional attains its minimum over each compact preimage $K_R = \psi^{-1}(]0, R])$, from which we obtain a sequence of minimizing functional portfolios $(\hat{\pi}_R)_{R \in \mathbb{N}}$ defined by

$\hat{\pi}_R = \arg\min_{\pi \in K_R} \mathcal{P}(\pi)$, for every $R \in \mathbb{N}$. Similarly as before, the performance of this sequence converges to the infimum of $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ over $\mathcal{FG}$, if $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ is bounded from below. Moreover, the performance functionals of Proposition 7.5 can be transfered into this path-dependent setting.

Proposition 7.8. *The following maps are performance functionals:*

(i) $\mathcal{P}^H(\pi) = \int_0^T l\left(Z_\pi(t) - z_t\right) dt$, *for some other path* $z : [0,T] \to \mathbb{R}$ *and for some lower semi-continuous function* $l : \mathbb{R} \to \mathbb{R}$.

(ii) $\mathcal{P}^R(\pi) = -\left(\log\left(\frac{Z_\pi(T)}{Z_\eta(T)}\right) - \log\left(\frac{Z_\pi(0)}{Z_\eta(0)}\right)\right)$, *for some other portfolio* $\eta : [0,T] \to \mathbb{R}^d$.

(iii) $\mathcal{P}^V(\pi) = \int_0^T \left(\pi(t) - \eta(t)\right)^\top \sigma(t) \left(\pi(t) - \eta(t)\right) dt$, *for some other portfolio* $\eta : [0,T] \to \mathbb{R}^d$.

In order to conclude this section, we are interested if an optimal path-dependent portfolio can be learned with the help of functional neural networks. We therefore show that the performance of the optimal functional neural network-generated portfolio converges to the performance of the optimal portfolio over all path-dependent portfolios in $\mathcal{FG}$, as the network connectivity increases.

Theorem 7.9. *Let* $\mathcal{P} : \mathcal{FG} \to \mathbb{R}$ *be a performance functional and assume that* (BV_κ) *holds true. Then*

$$\inf_{\pi \in \mathcal{FG}} \mathcal{P}(\pi) = \lim_{M \to \infty} \inf_{\Phi \in \mathcal{FN}_M} \mathcal{P}(\pi^{\exp(\Phi)}).$$

Proof. The proof follows along the lines of the proof for Theorem 7.6. □

Theorem 7.9 gives the theoretical justification to approximate an optimal path-dependent portfolios by functional neural networks. By increasing the network size, the performance of the optimal learned portfolio should converge to the performance of the optimal portfolio over $\mathcal{FG}$.

7.4 Numerical Examples

In this section, we illustrate in numerical examples how optimal portfolios can be learned within SPT, where we first look at randomized functional neural network to reduce the training effort. Finally, we apply the results to recover the generating function from the portfolio's wealth and try to outperform the S&P500.

7.4.1 Recovering Generating Function

In a first numerical example, we show that a generating function can be recovered from only one given historical trajectory of the value process of an investor. Assume that the market consists of $d = 3$ assets and suppose that the vector of logarithmic capitalizations X is given by a sample path of an Ito process, i.e. $\log X(t) = \log X_0 + \int_0^t \gamma(t)dt + \int_0^t \xi(t)dW(t)$, for $t \in [0,T]$, with initial value $X_0 = (100, 100, 100)^\top$, where the drift and volatility are constant throughout time and given by

$$\gamma(t) = \begin{pmatrix} 0.563 \\ 0.681 \\ 0.484 \end{pmatrix}, \qquad \xi(t) = \begin{pmatrix} 0.779 & 0 & 0 \\ 0.201 & 0.519 & 0 \\ 0.174 & 0.199 & 0.592 \end{pmatrix},$$

for all $t \in [0,T]$. In this framework, we use the generating functions described in Example 5.19(iii) and Example 5.19(v) to generate an entropy-weighted portfolio π^E and diversity-weighted portfolio π^D with their corresponding wealth Z_{π^E} and Z_{π^D}, respectively. For these two paths $Z_{\pi^E}, Z_{\pi^D} : [0,T] \to \mathbb{R}$, we learn two neural networks $\Phi^E, \Phi^D \in \mathcal{NN}_M$ such that the wealth of $\pi^{\exp(\Phi^E)}$ and $\pi^{\exp(\Phi^D)}$ should be close to the true wealth Z_{π^E} and Z_{π^D}, respectively. More precisely, we use the performance functional $\mathcal{P}^H$ with $l(x) = x^2$ from Proposition 7.5 used for hedging and try to minimize

$$\Phi \mapsto \mathcal{P}^H\left(\pi^{\exp(\Phi)}\right) = \int_0^T \left(Z_{\pi^{\exp(\Phi)}}(t) - Z_{\pi^{D/E}}(t)\right)^2 dt$$

among all neural networks $\Phi \in \mathcal{NN}_M$, for some fixed $M \in \mathbb{N}$. Hereby, we use actually two performance functional, where $Z_{\pi^{D/E}}$ is either computed with (6.2.1) or as Föllmer integral, i.e.

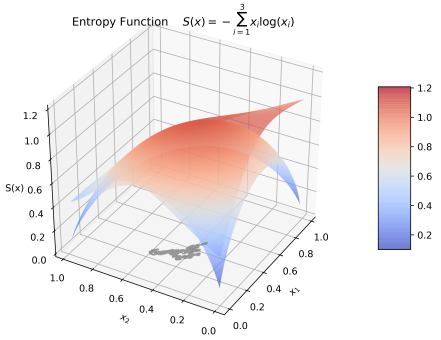
$$\begin{aligned}
\log Z_\pi(t) &= \log S(\mu(t)) + \Theta(t) + \log Z_\mu(t), \\
\log Z_\pi(t) &= \sum_{i=1}^d \int_0^t D_i \log S(\mu(s)) d\mu_i(s) \\
&\quad - \frac{1}{2} \sum_{i,j=1}^d \int_0^t D_i \log S(\mu(s)) D_j \log S(\mu(s)) \mu_i(s)\mu_j(s)\tau_{ij}(s)dt \\
&\quad + \log Z_\mu(t),
\end{aligned}$$

which are then combined to improve the approximation of the generating function by taking into account its gradient (in the second expression) and its Hessian (in the first expression with the drift Θ).

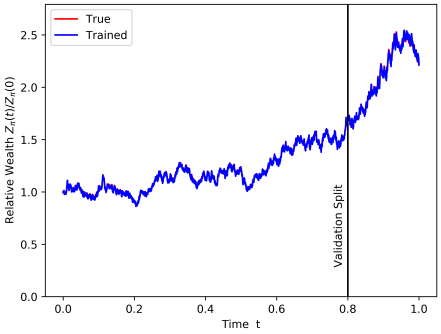
The algorithm has been implemented in `Python`, using `Tensorflow` to build

the neural networks of size $L = (3, 10, 1)$, i.e. neural networks with one hidden layer of dimension 10. For network training, the Adam algorithm (see [KB15] and [GBC16, Chapter 8.5.3]) is applied with a learning rate of 10^{-4} over 10^5 epochs. In addition, the market trajectory is split up into the first 80% time points used for training and the last 20% used for testing. The results for recovering the Entropy and Diversity function from one trajectory of the corresponding value process are displayed in Figure 7.1.

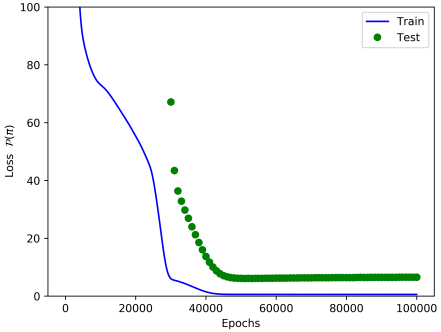
Figure 7.1 shows that the relatively small neural networks are indeed able to recover the shapes of the generating functions, from which the portfolios were generated. It seems that the market trajectory has a significant influence on how the generating function is learned, and the generating function is approximated better in the neighborhood of the market weights (data points), whereas the concavity of the generating function is less recaptured to the edges of Δ^d. Therefore, the path of the market weights must be enough volatile in order to reconstruct the generating function sufficiently accurate over Δ^d.



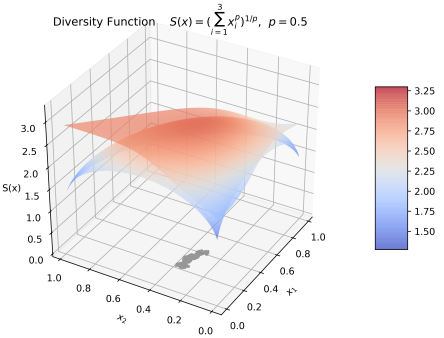
(a) Generating Function



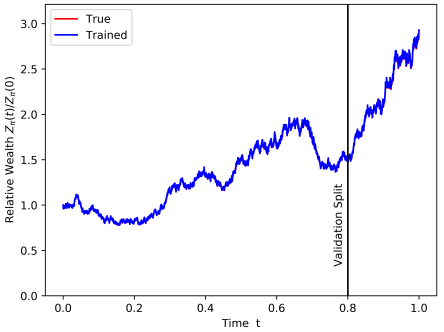
(b) Value Process



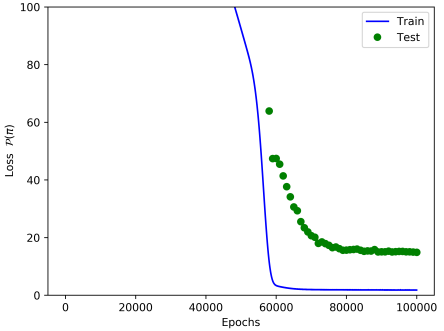
(c) Learning Performance



(d) Generating Function



(e) Value Process



(f) Learning Performance

Figure 7.1: Recovering the generating function. Left panel: Theoretical generating function and its approximating neural network of size $L = (3, 10, 1)$, where the shaded path in the x_1-x_2-plane represents the path of the market weights; Middle panel: Comparision of the wealth of $\pi^{\exp(\Phi)}$ and $\pi^{D/E}$, where the vertical line shows the validation split; Right panel: Loss during the training procedure on training and test set.

7.4.2 Outperforming the S&P500

In a second numerical example, we try to learn a relative arbitrage strategy in the market given by the universe of the S&P500, accounting for changes of constituents by setting the corresponding market weight to zero. In this setting, [KSV16] showed that learning an optimal portfolio with the help of Bayesian techniques consistently outperforms the market portfolio and other functionally generated portfolios presented in Example 5.19. We show in the following that our approach of learning an optimal path-dependent portfolio with functional neural networks also outperforms these results.

The data for stock prices and numbers of outstanding shares are downloaded from the CRSP databases of `Wharton Research Data Services`, and we consider the time period from 1st January 1992 to 31 December 2014, where the relative volatility matrix τ is estimated by the squared logarithmic increments of μ and using a rolling time window of 60 days. In this setting, we compare the following portfolios:

- the market portfolio μ whose value coincides with the S&P500 up to the divisor, denoted by MPF;
- the equally-weighted portfolio of Example 5.19(v), denoted by EWP;
- the diversity-weighted portfolio of Example 5.19(vi), where the exponent p is learned
  - by maximizing the excess return relative to EWP in [KSV16, Eq. 19], denoted by DWP*,
  - with the Markov Chain Monte Carlo (MCMC) method in [KSV16, Eq. 24-26], denoted by DWP;
- the conventional neural network approach from Section 7.2, to obtain a non-path-dependent optimal portfolio with respect to the performance functional $\mathcal{P}(\pi) = -\int_0^T \log\left(\frac{Z_\pi(s)/Z_\pi(0)}{Z_\mu(s)/Z_\mu(0)}\right) ds$, denoted by CNN;
- the functional neural networks approach from Section 7.3 to obtain a path-dependent optimal portfolio with respect to the performance functional $\mathcal{P}(\pi) = -\int_0^T \log\left(\frac{Z_\pi(t)/Z_\pi(0)}{Z_\mu(t)/Z_\mu(0)}\right) dt$, denoted by FNN.

Hereby, the excess return relative to the equally-weighted portfolio is defined as the performance functional

$$\mathcal{P}(\pi) = \prod_{t=1}^{T}\left(1 + \sum_{i=1}^{d} \pi_i(t) r_i(t)\right) - \prod_{t=1}^{T}\left(1 + \sum_{i=1}^{d} \eta_i(t) r_i(t)\right),$$

where η is the equally-weighted portfolio and $r_i(t) = \frac{X_i(t)-X_i(t-1)}{X_i(t-1)}$ denotes the return of the ith asset at time t. This performance functional is used for the other methods of [KSV16] as well. In addition, we learn symmetric (functional) neural networks to ensure enough generalization capacity, i.e. we simplify them to

$$\Phi(x) = \sum_{i=1}^{d} \sum_{k=1}^{N} \beta_k \rho \left(A_k x_i + b_k \right),$$

$$\Phi(t, Y) = \sum_{i=1}^{d} \sum_{k=1}^{N} \beta_k \rho \left(\int_0^t \varphi_{ki}(u) Y_i(u) du + A_k Y_i(t) + B_k t + b_k \right),$$

where $\beta, A, B, b \in \mathbb{R}^N$ and $\varphi_{ki} \in \mathcal{NN}^{\rho}_{1,\infty}$, with $N = 20$, such that $\Phi(x_1, ..., x_d) = \Phi(x_{\sigma(1)}, ..., x_{\sigma(d)})$ and $\Phi(t, (Y_1, ..., Y_d)) = \Phi(t, (Y_{\sigma(1)}, ..., Y_{\sigma(d)}))$, for all permutations $\sigma : \{1, ..., d\} \to \{1, ..., d\}$. The algorithms to learn the parameter p and the (functional) neural network have been implemented in `Python`, where the package `Tensorflow` is used to train the (functional) neural networks and the Adam algorithm is applied with a learning rate of 10^{-4} over 10^4 epochs (see [KB15] and [GBC16, Chapter 8.5.3]).

In the first experiment, we test if the learned optimal investment strategies described above indeed outperform the market portfolio, where we use 10 consecutive years as training data and test the performance on the following 5 years. Beginning on 1st January 1992 with the first data set, we roll both training and test set by one year, which yields a total of eight pairs of training and test sets. The results are listed in Table 7.1, which shows in particular that all learned strategies indeed outperform the benchmark portfolios, the market portfolio given as S&P500 and the equally-weighted portfolio, in-sample and out-of-sample.

Portfolio	In-Sample Return (in %)	Out-of-Sample Return (in %)
MPF	5.146 ± 3.849	1.598 ± 5.957
EWP	7.771 ± 2.721	7.546 ± 5.566
DWP*	6.968 ± 1.172	8.801 ± 2.411
DWP	6.931 ± 1.150	8.732 ± 2.387
CNN	9.921 ± 1.843	9.972 ± 5.243
FNN	9.694 ± 2.352	12.602 ± 4.342

Table 7.1: Outperforming the S&P500: Results of the first example to test whether the learned portfolios considerably outperform the value of the S&P500 and the other functionally generated portfolios. The in-sample and out-of-sample return are average yearly returns in percentage plus/minus two standard errors.

In the second experiment, we show that the learned strategies do not perform worse than the benchmark portfolios during the financial crisis. Hereby,

we use the first 15 years starting from 1st January 1992 to 31th December 2005 as training set and test the performance of the learned strategies on the test set ranging from 1st January 2006 to 31th December 2014, which includes in particular the financial crisis at the end of year 2008. The results of this experiment are listed in Table 7.2 and plotted in Figure 7.2. We observe in particular that all learned portfolios outperform the benchmark portfolios (market portfolio and equally-weighted portfolio) in-sample and out-of-sample. Figure 7.2 depicts the wealth of the learned portfolios and shows that they are robust against the financial crisis at the end of the year 2008.

Portfolio	In-Sample Return (in %)	Out-of-Sample Return (in %)
MPF	11.336 ± 8.616	7.488 ± 14.002
EWP	18.845 ± 8.373	14.692 ± 16.080
DWP*	26.261 ± 9.294	21.285 ± 17.956
DWP	26.135 ± 9.269	21.169 ± 17.926
CNN	76.716 ± 25.830	21.970 ± 14.013
FNN	57.739 ± 17.827	25.307 ± 14.011

Table 7.2: Outperforming the S&P500: Results of the second experiment to test whether the learned investment strategies are robust with respect to the financial crisis at the end of year 2008. The in-sample and out-of-sample return are average yearly returns in percentage plus/minus two standard errors.

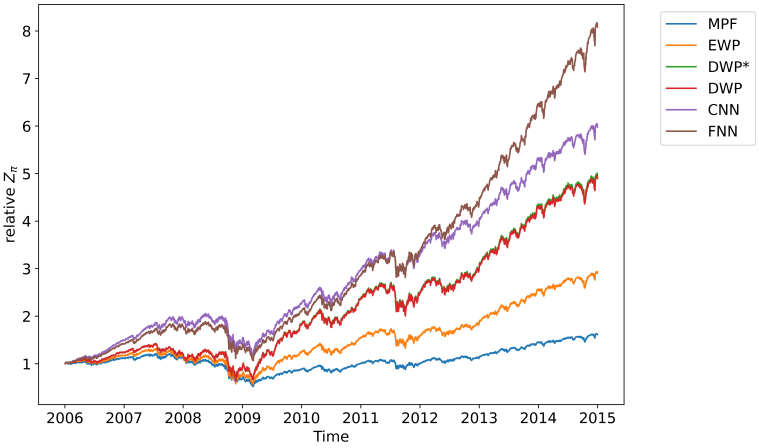


Figure 7.2: Outperforming the S&P500: Wealth of the learned investment strategies (see legend) over the test period 2006-2014 compared to the market capitalization of the S&P500, where the performances of DWP* and DWP are very similar.

Part III

Option Pricing with Abstract Neural Networks

Chapter 8

Option Pricing in a Financial Market

The problem of pricing and hedging portfolios of derivatives is a crucial task for portfolio managers in the financial industry. Despite the enormous progress in research over the last decades, there are only a few mathematical solutions under rather idealized assumptions such as completeness of the market, which means that every financial derivative can be perfectly replicated. The most prominent model herefore is Dupire's local volatility model introduced in [Dup94], which is a complete market model such that pricing becomes linear. However, real markets are often incomplete such that the pricing operator is a non-linear function, which can be expressed as supremum or infimum of the expected discounted payoff over so-called equivalent martingale measures. Because machine learning techniques are well-suited to solve such non-linear problems, we use the universal approximation result in Theorem 3.27 to learn the pricing operator.

In this chapter, we introduce the main ingredients of the financial market, where we assume that the stock price process follows an Ito diffusion. Subsequently, we define weak derivatives and introduce Sobolev spaces, which are needed to describe solutions of partial differential equations, which in turn can be related to the value function of a financial derivative with the help of the Feynman-Kac formula.

8.1 Market Model

In the following, we shall always consider a financial market model with finite time horizon $T > 0$, whose vector of price processes $X = (X(t))_{t\in[0,T]}$ is assumed to take values in a subset $E \subset \mathbb{R}^d$ and is defined on the filtered probability space $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{P})$, where the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ is supposed to satisfy the usual conditions, i.e. $\mathcal{F}_0$ is trivial, $\mathcal{A}$ is $\mathbb{P}$-complete and the filtration

is right-continuous such that $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$.

For some m-dimensional Brownian motion $B = (B(t))_{t\in[0,T]}$, we assume that X satisfies the stochastic differential equation (SDE)

$$dX(t) = b(t, X(t))dt + \xi(t, X(t))dB(t), \tag{8.1.1}$$

for $t \in [0, T]$, and some initial value $X(0) = x_0 \in E$, where $b : [0, T] \times E \to \mathbb{R}^d$ is the *drift function* and $\xi : [0, T] \times E \to \mathbb{R}^{d\times m}$ is the *volatility function*. If the two functions satisfy the uniform Lipschitz and linear growth condition, we have the following existence and uniqueness result for (8.1.1).

Theorem 8.1 ([Øks03, Theorem 5.2.1])**.** *Assume that there exists some constant $C > 0$ such that*

$$\begin{aligned} \|b(t,x) - b(t,y)\| + \|\xi(t,x) - \xi(t,y)\| &\leqslant C\|x - y\|, \\ \|b(t,x)\| + \|\xi(t,x)\| &\leqslant C\left(1 + \|x\|\right), \end{aligned} \tag{8.1.2}$$

for all $t \in [0, T]$ and $x, y \in E$. Then, (8.1.1) *has a strong solution satisfying $\mathbb{E}\left[\sup_{t\in[0,T]} \|X(t)\|^2\right] < \infty$.*

In the time-homogeneous case with $b(t, x) = b(x)$ and $\xi(t, x) = \xi(x)$, we observe that the Lipschitz condition implies the linear growth condition in (8.1.2). From now on, we shall always assume that the functions $b : [0, T]\times E \to \mathbb{R}^d$ and $\xi : [0, T] \times E \to \mathbb{R}^{d\times m}$ satisfy the Lipschitz and linear growth condition in (8.1.2) such that there exists a strong solution $X = (X(t))_{t\in[0,T]}$ satisfying the SDE in (8.1.1).

In addition, we assume that there is an additional money market account representing the value of cash, which satisfies the SDE $dX_0(t) = r(t)X_0(t)dt$ with $X_0(0) = 1$, such that

$$X_0(t) = \exp\left(\int_0^t r(s)ds\right),$$

for $t \in [0, T]$, where $r = (r(t))_{t\in[0,T]}$ is an $\mathbb{F}$-adapted process representing the short term interest rate. From this, we define for $0 \leqslant s \leqslant t \leqslant T$ the discount factor $D_{st} = X_0(s)/X_0(t) = \exp\left(-\int_s^t r(u)du\right)$ and denote by $\widetilde{X} = (\widetilde{X}(t))_{t\in[0,T]}$ the discounted price processes given by $\widetilde{X}(t) = D_{0t}X(t)$, for all $t \in [0, T]$.

In order to trade in the market, we define a *trading strategy* as $\mathbb{F}$-predictable, $(d + 1)$-dimensional stochastic process $\theta = (\theta(t))_{t\in[0,T]}$ such that $\theta_0(t)$ denotes the amount invested in the risk-free asset X_0 at time $t \in [0, T]$, and $\theta_i(t)$ the number of shares held from the risky asset X_i at time $t \in [0, T]$, for $i = 1, ..., d$.

Then, the value of the trading strategy $V_\theta = (V_\theta(t))_{t\in[0,T]}$ is given by

$$V_\theta(t) = \sum_{i=0}^{d} \theta_i(t) X_i(t),$$

for all $t \in [0,T]$. In addition, we assume that a trading strategy θ is self-financing in the sense that the change in its value is only affected by trading in the market and no money is injected or removed from the trading strategy θ. This means for the time between t and $t + dt$ that

$$dV_\theta(t) = \sum_{i=0}^{d} \theta_i(t) dX_i(t),$$

for all $t \in [0,T]$, which also holds true for the discounted value process $\widetilde{V}_\theta = (\widetilde{V}_\theta(t))_{t\in[0,T]}$ such that

$$d\widetilde{V}_\theta(t) = \sum_{i=0}^{d} \theta_i(t) d\widetilde{X}_i(t) = \sum_{i=1}^{d} \theta_i(t) d\widetilde{X}_i(t), \tag{8.1.3}$$

for all $t \in [0,T]$, where we have used that $\widetilde{X}_0(t) = 1$ and $d\widetilde{X}_i(t) = D_{0t}(dX_i(t) - r(t)X_i(t)dt)$, for all $i = 1, ..., d$ and $t \in [0,T]$. Hence, the discounted value process $\widetilde{V}_\theta$ satisfies

$$\widetilde{V}_\theta(t) = V_\theta(0) + \sum_{i=1}^{d} \int_0^t \theta_i(s) d\widetilde{X}_i(s),$$

for all $t \in [0,T]$, where $V_\theta(0)$ is the initial endowment of the trading strategy.

Definition 8.2. *An $\mathbb{F}$-predictable, $(d+1)$-dimensional process $\theta = (\theta(t))_{t\in[0,T]}$ is called an* admissible trading strategy *if θ is self-financing and $\widetilde{V}_\theta(t)$ is a.s. bounded from below, for all $t \in [0,T]$, i.e. there exists $C > 0$ such that $\mathbb{P}\left[\forall t \in [0,T] : \widetilde{V}_\theta(t) \geqslant -C\right] = 1$. We denote the set of admissible trading strategies by Θ_{adm}.*

In the following, we briefly introduce the concept of arbitrage strategies, which is a self-financing trading strategy requiring zero initial endowment but admits a positive termial gain with positive probability.

Definition 8.3. *A self-financing, admissible trading strategy* $\theta = (\theta(t))_{t\in[0,T]}$ *is an* arbitrage strategy *if* $\widetilde{V}_\theta(0) = 0$, $\widetilde{V}_\theta(T) \geqslant 0$, *a.s., and* $\mathbb{P}\left[\widetilde{V}_\theta(T) > 0\right] > 0$.

In Finance, arbitrageurs are special types of traders who immediately exploit arbitrage opportunities so that they quickly disappear and the market remains arbitrage-free. To give a sufficient condition for an arbitrage-free market, we introduce so-called equivalent martingale measuers (EMMs), where a measure $\mathbb{Q}$ on $(\Omega, \mathcal{A})$ is called *equivalent* to $\mathbb{P}$, denoted by $\mathbb{Q} \sim \mathbb{P}$, if and only if $\mathbb{Q}[A] = 0 \Leftrightarrow \mathbb{P}[A] = 0$, for all $A \in \mathcal{A}$.

Definition 8.4. *A probability measure* $\mathbb{Q} \sim \mathbb{P}$ *defined on* $(\Omega, \mathcal{A})$ *is called an* equivalent martingale measure (EMM) (for X) *if the discounted price processes* $\widetilde{X}$ *is a local martingale under* $\mathbb{Q}$.

Theorem 8.5 (First Fundamental Theorem of Asset Pricing, simplified [DS94, Corollary 1.2])**.** *If there exists an equivalent martingale measure* $\mathbb{Q} \sim \mathbb{P}$, *then the market admits no arbitrage opportunity.*

Proof. Assume by contradiction that there exists an arbitrage opportunity $\theta \in \Theta_{\text{adm}}$. By (8.1.3) and since θ is admissible, $\widetilde{V}_\theta$ is a local martingale under $\mathbb{Q}$, which is bounded from below. Hence, it is by Fatou's Lemma a supermartingale such that $\mathbb{E}^{\mathbb{Q}}\left[\widetilde{V}_\theta(T)\right] \leqslant \mathbb{E}^{\mathbb{Q}}\left[\widetilde{V}_\theta(0)\right] = 0$. However, because $\widetilde{V}_\theta(T) \geqslant 0$, $\mathbb{P}$-a.s., and therefore also $\mathbb{Q}$-a.s., it follows that $\widetilde{V}_\theta(T) = 0$, $\mathbb{Q}$-a.s., and therefore also $\mathbb{P}$-a.s. This however contradicts the assumption that $\mathbb{P}\left[\widetilde{V}_\theta(T) > 0\right] > 0$. □

A more general version of the first fundamental theorem of asset pricing can be found in [DS94, Corollary 1.2], where an equivalent condition to the existence of an equivalent martingale measure $\mathbb{Q} \sim \mathbb{P}$ is stated, which is called *No Free Lunch With Vanishing Risk (NFLVR)*. For the rest of this chapter, we shall always assume the existence of an equivalent martingale measure $\mathbb{Q} \sim \mathbb{P}$ such that the market is arbitrage-free.

In the following, we consider the pricing problem of a financial derivative, which is represented by a random variable $C \in L^0(\mathcal{F}_T)$, where $L^0(\mathcal{F}_t)$ denotes the space of $\mathcal{F}_t$-measurable random variables. We define the buyer's (and seller's) price at time $t \in [0, T]$ as the largest (lowest) price such that the buyer's (seller's) trading strategy $\theta \in \Theta_{\text{adm}}$ sub-hedges (super-hedges) the financial

derivative, i.e.

$$\pi_t^b(C) = \operatorname{ess\,sup}\Bigg\{ Z \in L^0(\mathcal{F}_t);\, \exists \theta \in \Theta_{\mathrm{adm}} : \\ - D_{0t}Z + \sum_{i=1}^d \int_t^T \theta_i(s) d\widetilde{X}_i(s) \geqslant -D_{0T}C,\, \mathbb{P}\text{-a.s.} \Bigg\},$$

$$\pi_t^s(C) = \operatorname{ess\,inf}\Bigg\{ Z \in L^0(\mathcal{F}_t);\, \exists \theta \in \Theta_{\mathrm{adm}} : \\ D_{0t}Z + \sum_{i=1}^d \int_t^T \theta_i(s) d\widetilde{X}_i(s) \geqslant D_{0T}C,\, \mathbb{P}\text{-a.s.} \Bigg\},$$

where $U = \operatorname{ess\,sup}(\mathcal{Z})$ for a subset $\mathcal{Z} \subset L^0(\mathcal{F}_t)$ if and only if $U \geqslant Z$, $\mathbb{P}$-a.s., for all $Z \in \mathcal{Z}$, and for every $V \in L^0(\mathcal{F}_t)$, we have $U \leqslant V$, $\mathbb{P}$-a.s. Then, we define $\operatorname{ess\,inf}(\mathcal{Z}) = -\operatorname{ess\,sup}(-\mathcal{Z})$ such that $\pi_t^s(C) = -\pi_t^b(-C)$.

Proposition 8.6 ([GH14, Theorem 1.1])**.** *Assume that there exists an equivalent martingale measure $\mathbb{Q} \sim \mathbb{P}$. Then, it holds for every $C \in L^0(\mathcal{F}_T)$ that $\pi_t^b(C) \leqslant \mathbb{E}^{\mathbb{Q}}\left[D_{tT}C|\mathcal{F}_t\right] \leqslant \pi_t^s(C)$, $\mathbb{P}$-a.s., for all $t \in [0,T]$.*

Proof. By definition of $\pi_t^b(C)$, we assume that there exists $\theta \in \Theta_{\mathrm{adm}}$ with $D_{0t}Z + \sum_{i=1}^d \int_t^T \theta_i(s) d\widetilde{X}_i(s) \leqslant D_{0T}C$, $\mathbb{P}$-a.s., and thus also $\mathbb{Q}$-a.s. Since $t \mapsto \sum_{i=1}^d \int_0^t \theta_i(s) d\widetilde{X}_i(s)$ is a local martingale under $\mathbb{Q}$, it is by Fatou's lemma a supermartingale under $\mathbb{Q}$ such that $\mathbb{E}^{\mathbb{Q}}\left[\sum_{i=1}^d \int_t^T \theta_i(s) d\widetilde{X}_i(s) \middle| \mathcal{F}_t\right] \leqslant 0$, $\mathbb{Q}$-a.s., and therefore $Z \leqslant \mathbb{E}^{\mathbb{Q}}\left[D_{tT}C|\mathcal{F}_t\right]$, $\mathbb{Q}$-a.s. Taking the essential supremum over all $Z \in L^0(\mathcal{F}_t)$, we obtain $\pi_t^b(C) \leqslant \mathbb{E}^{\mathbb{Q}}\left[D_{tT}C|\mathcal{F}_t\right]$, $\mathbb{Q}$-a.s., and thus $\mathbb{P}$-a.s., where the other inequality follows by similar arguments. □

Hence, the buyer's price $\pi_t^b(C)$ and seller's price $\pi_t^s(C)$ can be seen as no-arbitrage price bounds, where we say that $C \in L^0(\mathcal{F}_T)$ is an *attainable payoff (at time 0)* if there exists an admissible trading strategy $\theta \in \Theta_{\mathrm{adm}}$ and a real number $z \in \mathbb{R}$ such that $z + \sum_{i=1}^d \int_0^T \theta_i(t) d\widetilde{X}_i(t) = D_{0T}C$, $\mathbb{P}$-a.s., and $t \mapsto \sum_{i=1}^d \int_0^t \theta_i(t) d\widetilde{X}_i(t)$ is a true martingale under $\mathbb{Q}$. In this case, a market is called *complete* if every payoff $C \in L^0(\mathcal{F}_T)$ is attainable at initial time. Then, the buyer's price $\pi_t^b(C)$ and seller's price $\pi_t^s(C)$ coincide.

Proposition 8.7 ([GH14, Theorem 1.2])**.** *In a complete market, it holds for every $C \in L^0(\mathcal{F}_T)$ and equivalent martingale measure $\mathbb{Q} \sim \mathbb{P}$ that $\pi_t^b(C) = \mathbb{E}^{\mathbb{Q}}\left[D_{tT}C|\mathcal{F}_t\right] = \pi_t^s(C)$, $\mathbb{P}$-a.s., for all $t \in [0,T]$.*

Proof. Since the market is complete, we find for fixed $C \in L^0(\mathcal{F}_T)$ some $z \in \mathbb{R}$ and $\theta \in \Theta_{\mathrm{adm}}$ such that $z + \sum_{i=1}^d \int_0^T \theta_i(t) d\widetilde{X}_i(t) = D_{0T}C$, $\mathbb{P}$-a.s., and therefore also $\mathbb{Q}$-a.s. Then, it follows for every $t \in [0,T]$ that

$$D_{0t}Z_t + \sum_{i=1}^d \int_t^T \theta_i(s) d\widetilde{X}_i(s) = D_{0T}C, \quad \mathbb{P}\text{-a.s.},$$

where $Z_t = D_{0t}^{-1}\left(z + \sum_{i=1}^d \int_0^t \theta_i(s) d\widetilde{X}_i(s)\right)$ is $\mathcal{F}_t$-measurable. As the process $t \mapsto \sum_{i=1}^d \int_0^t \theta_i(t) d\widetilde{X}_i(t)$ is a true martingale under $\mathbb{Q}$, it follows that $Z_t = \mathbb{E}^{\mathbb{Q}}[D_{tT}C|\mathcal{F}_t]$, $\mathbb{Q}$-a.s., and thus $\mathbb{P}$-a.s. By taking the essential supremum over $Z_t \in L^0(\mathcal{F}_t)$, we have $\pi_t^b(C) \geqslant \mathbb{E}^{\mathbb{Q}}[D_{tT}C|\mathcal{F}_t]$, $\mathbb{P}$-a.s. Using Proposition 8.6, we obtain $\pi_t^b(C) = \mathbb{E}^{\mathbb{Q}}[D_{tT}C|\mathcal{F}_t]$, $\mathbb{P}$-a.s., which follows by similar arguments also for $\mathbb{E}^{\mathbb{Q}}[D_{tT}C|\mathcal{F}_t] = \pi_t^s(C)$, $\mathbb{P}$-a.s., for all $t \in [0,T]$. □

While in complete markets, every payoff $C \in L^0(\mathcal{F}_T)$ is attainable and the price is given as the initial value of the buyer's or seller's trading strategy, perfect replication is no longer possible in incomplete markets. However, the buyer's and seller's price can be expressed by using the equivalent martingale measures.

Proposition 8.8 ([Kra96, Theorem 3.2+3.3])**.** *For every $C \in L^0(\mathcal{F}_T)$, it holds that*

$$\pi_t^b(C) = \operatorname*{ess\,inf}_{\mathbb{Q}\, EMM} \mathbb{E}^{\mathbb{Q}}[D_{tT}C|\mathcal{F}_t], \quad \text{and} \quad \pi_t^s(C) = \operatorname*{ess\,sup}_{\mathbb{Q}\, EMM} \mathbb{E}^{\mathbb{Q}}[D_{tT}C|\mathcal{F}_t], \quad \mathbb{P}\text{-a.s.}$$

Theorem 8.9 (Second Fundamental Theorem of Asset Pricing, [GH14, Corollary 1.1])**.** *A market is complete if and only if there exists a unique equivalent martingale measure $\mathbb{Q} \sim \mathbb{P}$.*

Proof. If the market is complete and $\mathbb{Q}_0 \sim \mathbb{P}$ denotes an equivalent martingale measure, it follows for every $C \in L^0(\mathcal{F}_T)$ from Proposition 8.7 and Proposition 8.8 that $\pi_0^b(C) = \pi_0^s(C) = \mathbb{E}^{\mathbb{Q}_0}[D_{0T}C] = \inf_{\mathbb{Q}\,\mathrm{EMM}} \mathbb{E}^{\mathbb{Q}}[D_{0T}C] = \sup_{\mathbb{Q}\,\mathrm{EMM}} \mathbb{E}^{\mathbb{Q}}[D_{0T}C]$, which implies $\mathbb{E}^{\mathbb{Q}}[D_{0T}C] = \mathbb{E}^{\mathbb{Q}_0}[D_{0T}C]$, for all $C \in L^0(\mathcal{F}_T)$ and equivalent martingale measures $\mathbb{Q} \sim \mathbb{P}$, such that $\mathbb{Q}_0$ is unique. Conversely, let $\mathbb{Q}_0$ be the unique equivalent martingale measure, then it follows for some fixed $C \in L^0(\mathcal{F}_T)$ by Proposition 8.6 that $\pi_0^b(C) = \mathbb{E}^{\mathbb{Q}_0}[D_{0T}C] = \pi_0^s(C)$. Hence, there exists two admissible trading strategies $\theta^b, \theta^s \in \Theta_{\mathrm{adm}}$ such

that

$$\begin{aligned} -\mathbb{E}^{\mathbb{Q}_0}[D_{0T}C] + \sum_{i=1}^{d} \int_0^T \theta_i^b(s) d\widetilde{X}_i(s) &\geqslant -D_{0T}C, \\ \mathbb{E}^{\mathbb{Q}_0}[D_{0T}C] + \sum_{i=1}^{d} \int_0^T \theta_i^s(s) d\widetilde{X}_i(s) &\geqslant D_{0T}C, \end{aligned} \tag{8.1.4}$$

$\mathbb{P}$-a.s., and thus also $\mathbb{Q}_0$-a.s. Summing both inequalities, it follows that

$$\sum_{i=1}^{d} \int_0^T \left(\theta_i^b(s) - \theta_i^s(s)\right) d\widetilde{X}_i(s) \geqslant 0$$

$\mathbb{P}$-a.s., and as $\widetilde{X}$ is a $\mathbb{Q}_0$-local martingale, we have $\theta_i^b(s) = -\theta_i^s(s)$, $\mathbb{P}$-a.s., for all $i = 1, ..., d$ and $s \in [0, T]$. Using this for (8.1.4), we conclude that $\mathbb{E}^{\mathbb{Q}_0}[D_{0T}C] + \sum_{i=1}^{d} \int_0^T \theta_i^b(s) d\widetilde{X}_i(s) = D_{0T}C$, $\mathbb{P}$-a.s., such that $C \in L^0(\mathcal{F}_T)$ is attainable at time 0. Since $C \in L^0(\mathcal{F}_T)$ was chosen arbitrarily, the market is complete. □

There exists a simple construction of an equivalent martingale measure for the Ito diffusion X being a solution of the SDE (8.1.1). Every $\mathbb{F}$-adapted, d-dimensional process $\lambda = (\lambda(t))_{t\in[0,T]}$ satisfying

$$b(t, X(t)) - r(t)X(t) = \xi(t, X(t))\lambda(t), \tag{8.1.5}$$

$\mathbb{P}$-a.s., for all $t \in [0, T]$, induces an equivalent martingale measure $\mathbb{Q}_\lambda$, whose density $\frac{d\mathbb{Q}_\lambda}{d\mathbb{P}}$ is given by

$$\frac{d\mathbb{Q}_\lambda}{d\mathbb{P}} = \exp\left(-\int_0^T \lambda(t)^\top dB(t) - \frac{1}{2} \int_0^T \|\lambda(t)\|^2 dt \right).$$

Hence, completeness of a market can be characterized with the help of (8.1.5). If the market is complete, then the rank of $\xi(t, X(t))$ is a.s. equal to d (which implies $d \geqslant m$), whereas $d < m$ implies that the market is incomplete. Moreover, if $d = m$, the market is complete if and only if $\xi(t, X(t))$ is a.s. invertible.

8.2 Sobolev Spaces

In this section, we introduce Sobolev spaces being related to partial differential equations (PDEs), which can be used to describe many real-world phenomena by mathematical terms such as in Physics, where we refer to [Eva15] for the mathematical theory of PDEs. For an open subset $E \subset \mathbb{R}^d$, let $u : E \to \mathbb{R}$ be a

k-times differentiable function, whose partial derivatives $\partial^\alpha u$ exist, for all multi-indices $\alpha \in \mathbb{N}_k^d$, where $\mathbb{N}_k^d := \{\alpha \in \mathbb{N}^d;\, |\alpha| = \alpha_1 + ... + \alpha_d \leqslant k\}$. We collect the partial derivatives of order $l = 1, ..., k$ in the tensor $\nabla^l u(x) = (\partial^\alpha u(x))_{|\alpha|=l} \in (\mathbb{R}^d)^{\otimes l}$, where $(\mathbb{R}^d)^{\otimes l} = \mathbb{R}^d \otimes \cdots \otimes \mathbb{R}^d$. Then, an expression of the form

$$F\left(x, u(x), \nabla u(x), \nabla^2 u(x), ..., \nabla^k u(x)\right) = 0 \tag{8.2.1}$$

is called a *partial differential equation (PDE)* of order k, where the function

$$F : E \times \mathbb{R} \times \mathbb{R}^d \times ... \times (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}$$

is given and the equation (8.2.1) should be solved for the unknown function $u : E \to \mathbb{R}$. In addition, a partial differential equation is linear, if the function $u : E \to \mathbb{R}$ and its derivatives appear linearly in the PDE, or equivalently if $F : E \times \mathbb{R} \times \mathbb{R}^d \times ... \times (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}$ is linear in all its inputs except for the first.

In the following, we will mainly consider second order PDEs with $d + 1$ variables, where the 0th variable $x_0 = t$ denotes the time variable and the others $x_1, ..., x_d$ represent the space variables. In the case of a linear second order PDE, the function $F : E \times \mathbb{R} \times \mathbb{R}^d \times (\mathbb{R}^d)^{\otimes 2} \to \mathbb{R}$ is in general of the form

$$\begin{aligned} F\left(x, u(x), \nabla u(x), \nabla^2 u(x)\right) = &- \sum_{i,j=0}^{d} a_{ij}(x)\partial_{ij} u(x) + \sum_{i=0}^{d} b_i(x)\partial_i u(x) \\ &+ c(x)u(x) - f(x), \end{aligned} \tag{8.2.2}$$

for all $x \in \mathbb{R}^{d+1}$, where $b : E \to \mathbb{R}^{d+1}$ and $f : E \to \mathbb{R}$ are functions, and where $A : E \to \mathbb{R}^{(d+1)\times(d+1)}$ is a matrix-valued function such that $A(x) = (a_{ij}(x))_{i,j=0,...,d}$ is symmetric with eigenvalues $\lambda_0(x) \leqslant ... \leqslant \lambda_d(x)$.

Definition 8.10. *We consider the linear second order PDE in* (8.2.2) *and introduce the following.*

(i) The PDE is elliptic *if for every $i = 0, ..., d$ and $x \in E$, $\lambda_i(x) \neq 0$ and $sgn(\lambda_0(x)) = ... = sgn(\lambda_d(x))$.*

(ii) The PDE is parabolic *if for every $x \in E$, it holds that $rank(A(x), b(x)) = d + 1$ and there exists a unique $j = 0, ..., d$ with $\lambda_j(x) = 0$.*

(iii) The PDE is hyperbolic *if for every $x \in E$, it holds that $\lambda_i(x) \neq 0$, for all $i = 0, ..., d$, and there exists a unique $j = 0, ..., d$ with $sgn(\lambda_j(x)) \neq sgn(\lambda_k(x))$, for all $k \in \{0, ..., j-1, j+1, ..., d\}$.*

Let $\Delta = \sum_{i=1}^d \partial_i^2$ be the Laplace operator, then the Poisson equation $\Delta u = f$ is elliptic, the heat equation $\partial_t u - \Delta u = f$ is parabolic, and the wave equation

$\partial_{tt}u - \Delta u = f$ is hyperbolic. In addition, the Black-Scholes PDE is a prominent parabolic PDE of Finance, where $E =]0,T[\times]0,\infty[$ and

$$\partial_t v(t,x) - \frac{1}{2}\sigma^2 x^2 \partial_{xx} v(t,x) - rx\partial_x v + rv(t,x) = 0, \tag{8.2.3}$$

for some constants $r \geqslant 0$ and $\sigma > 0$, representing the interest rate and volatility parameter. This PDE can be obtained by a suitable transformation from the heat equation. Indeed, let $u :]0,T[\times\mathbb{R} \to \mathbb{R}$ be a solution of the parabolic heat equation $\partial_s u - \partial_{yy} u = 0$, then applying the transformations $x = e^y$ and $t = 2\sigma^2 s$ and defining the function $v :]0,T[\times]0,\infty[\to \mathbb{R}$ by

$$v(t,x) = e^{\alpha y + \beta s} u(s,y), \quad \text{where} \quad \alpha = 1/2 - r/\sigma^2, \quad \beta = \left(-1/2 - r/\sigma^2\right)^2,$$

then $v :]0,T[\times]0,\infty[\to \mathbb{R}$ is a solution of the Black-Scholes PDE in (8.2.3).

For an open subset $E \subset \mathbb{R}^d$ and $u \in C^k(E)$, we observe by applying integration by parts with some test function $\eta \in C_c^\infty(E)$ that

$$\int_E \partial^\alpha u(x)\eta(x)dx = (-1)^{|\alpha|} \int_E u(x)\partial^\alpha \eta(x)dx,$$

for all $\alpha \in \mathbb{N}_k^d$, where $C_c^\infty(E)$ denotes the space of smooth functions with compact support properly contained in E. Using this idea, we call $g \in L^2(E)$ the *weak αth partial derivative* of $u : \overline{E} \to \mathbb{R}$ if it holds that

$$\int_E g(x)\eta(x)dx = (-1)^{|\alpha|} \int_E u(x)\partial^\alpha \eta(x)dx,$$

for all $\eta \in C_c^\infty(E)$. In particular, if $u \in C^k(E)$, the weak partial derivatives coincide with the classical partial derivatives $\partial^\alpha u$, where we use by a slight abuse of notation the same symbol $\partial^\alpha u$ to denote the weak αth partial derivative of u. Using these weak derivatives, we define the Sobolev space $W^{k,p}(E)$ as

$$W^{k,p}(E) = \left\{u \in L^p(E);\ \partial^\alpha u \in L^p(E),\ \text{for all } \alpha \in \mathbb{N}_k^d\right\},$$

which can be equipped with the norm $\|u\|_{W^{k,p}(E)}^p = \sum_{\alpha\in\mathbb{N}_k^d} \|\partial^\alpha u\|_{L^p(E)}^p = \sum_{\alpha\in\mathbb{N}_k^d} \int_E |\partial^\alpha u(x)|^p\, dx$ such that $W^{k,p}(E)$ becomes a Banach space, for any $1 \leqslant p < \infty$. In the special case of $p = 2$, we define $H^k(E) = W^{k,2}(E)$, where the norm is then induced by $\langle u, v\rangle_{H^k(E)} = \sum_{\alpha\in\mathbb{N}_k^d} \langle \partial^\alpha u, \partial^\alpha v\rangle_{L^2(E)}$.

There are different methods to solve a partial differential equation numerically, e.g. finite differences and finite elements for elliptic PDEs described in [Bra07], and for parabolic and hyperbolic PDEs in [LT05]. In addition, [HRSW13] gives an overview of these methods applied in Finance.

8.3 Value Function of Financial Option

In this section, we now connect the pricing problem of an European option with the solution of a particular partial differential equation. For this purpose, we assume that the drift $b : [0,T] \times E \to \mathbb{R}^d$ and volatility function $\xi : [0,T] \times E \to \mathbb{R}^{d\times m}$ satisfy the uniform Lipschitz and linear growth condition in (8.1.2). Moreover, the short term interest rate r is supposed to be of the form $r(t) = r(X(t))$, for some function $r : E \to \mathbb{R}$. In addition, we introduce the following differential operator, which is the so-called infinitesimal generator of the Ito diffusion X described in (8.1.1) and characterizes the law of the process.

Definition 8.11. *If X is a solution of the SDE* (8.1.1)*, the* infinitesimal generator $\mathcal{A} = (\mathcal{A}_t)_{t\in[0,T]} : C^2(E) \to C^2(E)$ *is a collection of differential operators defined by*

$$(\mathcal{A}_t f)(x) = \frac{1}{2} tr\big(\sigma(t,x)\nabla^2 f(x)\big) + b(t,x)^\top \nabla f(x),$$

for $t \in [0,T]$, $x \in E$ and $f \in C^2(E)$, where $\sigma : [0,T] \times E \to \mathbb{R}^{d\times d}$ is defined by $\sigma(t,x) = \xi(t,x)\xi(t,x)^\top$, for $t \in [0,T]$ and $x \in E$.

The infinitesimal generator of X can be used together with Ito's formula in Theorem 1.39 to show that particular processes constructed with some $f \in C^2(E)$ (resp. $f \in C^{1,2}([0,T] \times E)$) are martingales under $\mathbb{P}$.

Proposition 8.12 ([HRSW13, Proposition 4.1.1])**.** *Assume that $f \in C^2(E)$ has bounded derivatives. Then, the process M defined by*

$$M(t) = f(X(t)) - \int_0^t (\mathcal{A}_s f)(X(s))ds,$$

for $t \in [0,T]$, is a martingale with respect to $\mathbb{P}$ and $\mathbb{F}$.

Proof. By applying Ito's formula in Theorem 1.39 and inserting the linear SDE (8.1.1), it follows that

$$\begin{aligned} f(X(t)) &= f(X(0)) + \sum_{i=1}^d \int_0^t \partial_i f(X(s))dX_i(s) + \frac{1}{2}\sum_{i,j=1}^d \int_0^t \partial_{ij} f(X(s))d\langle X_i, X_j\rangle_s \\ &= f(X(0)) + \int_0^t (\mathcal{A}_s f)(X(s))ds + \underbrace{\int_0^t \nabla f(X(s))^\top \xi(s, X(s))dB(s)}_{=:Z(t)}, \end{aligned}$$

$\mathbb{P}$-a.s., for all $t \in [0,T]$. The conclusion follows if we can show that the process

Z defined above is a true martingale. Because the partial derivatives of f are bounded and $\xi : [0,T] \times E \to \mathbb{R}^{d\times m}$ satisfies the linear growth condition in (8.1.2) such that $\|\xi(u,x)\| \leqslant C^2\left(1+\|x\|^2\right)$, it follows by Ito's isometry that

$$\begin{aligned}\mathbb{E}\left[Z(t)^2\right] &= \mathbb{E}\left[\int_0^t \nabla f(X(s))^\top \sigma(s,X(s))\nabla f(X(s))ds\right] \\ &\leqslant C^2 \sup_{x\in E}\|\nabla f(x)\|^2 \mathbb{E}\left[\int_0^t \left(1+\|X(s)\|^2\right)ds\right] \\ &\leqslant C^2 T \sup_{x\in E}\|\nabla f(x)\|^2\left(1+\mathbb{E}\left[\sup_{s\in[0,T]}\|X(s)\|^2\right]\right) < \infty,\end{aligned}$$

for all $t \in [0,T]$. Hence, Z is square-integrable and by [Øks03, Theorem 3.2.1] therefore a martingale. □

Proposition 8.13 ([HRSW13, Proposition 4.1.3])**.** *Let $f \in C^{1,2}([0,T]\times E)$ be a function with bounded partial derivatives in $x_1, ..., x_d$, and let $r \in C^0(E)$ be bounded. Then, the process M defined by*

$$M(t) = e^{-\int_0^t r(X(s))ds} f(t,X(t)) - \int_0^t e^{-\int_0^s r(X(u))du}\left(\partial_t f + \mathcal{A}f - rf\right)(s,X(s))ds \tag{8.3.1}$$

is a martingale with respect to $\mathbb{P}$ and $\mathbb{F}$.

Proof. Let Y be the process $Y(t) = e^{-\int_0^t r(X(s))ds}$, for $t \in [0,T]$ such that $dY(t) = -r(X(t))Y(t)dB(t)$, for all $t \in [0,T]$. Then, it follows by applying Ito's formula on $Y(t)f(t,X(t))$ and inserting the SDE (8.1.1) that

$$\begin{aligned}Y(t)f(t,X(t)) &= Y(0)f(0,X(0)) \\ &\quad + \int_0^t Y(s)\left(\nabla f(s,X(s))^\top b(s,X(s)) + \frac{1}{2}\mathrm{tr}\left(\sigma(s,X(s))\nabla^2 f(s,X(s))\right)\right)ds \\ &\quad + \int_0^t \nabla f(s,X(s))^\top \xi(s,X(s))dB(s) + \int_0^t f(s,X(s))dY(s) \\ &= Y(0)f(0,X(0)) + \int_0^t Y(s)\left(\partial_t f + \mathcal{A}_s f - rf\right)(s,X(s))ds + Z(t),\end{aligned}$$

$\mathbb{P}$-a.s., for all $t \in [0,T]$, where $Z(t) = \int_0^t Y(s)\nabla f(s,X(s))^\top \xi(s,X(s))dB(s)$. As r is bounded, it follows by similar arguments as in Proposition 8.12 that Z is a martingale, from which we obtain the conclusion. □

In the following, we assume that the interest rate function $r \in C^0(E)$ is bounded and consider the infinitesimal generator $\mathcal{A}$ under an equivalent martingale measure such that $b(t,x) = r(x)x$, for all $(t,x) \in [0,T] \times E$. Hence, the discounted price process $t \mapsto \exp\left(-\int_0^t r(X(s))ds\right) X(t)$ is a martingale under $\mathbb{Q}$ such that we would like to compute the price of an European option with payoff function $g : E \to \mathbb{R}$, i.e.

$$u(t,x) = \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r(X(s))ds} g(X(T)) \middle| X(t) = x\right],$$

representing the discounted price of the option $g(X(T))$ at time t if the stock prices were $X(t) = x$. In this case, $u : [0,T] \times E \to \mathbb{R}$ can be seen as function of time $t \in [0,T]$ and initial value $x \in E$, for which we show that it is a solution of a particular PDE by using the infinitesimal generator of $X = (X(t))_{t\in[0,T]}$.

Theorem 8.14 (Feynman-Kac, [HRSW13, Theorem 4.1.4])**.** *For $J \subset [0,T]$ and $E \subset \mathbb{R}^d$ open, let $u \in C^{1,2}(J \times E) \cap C^0(\overline{J} \times E)$ with bounded partial derivatives in $x_1, ..., x_d$, and let $r \in C^0(E)$ be bounded. Then, $u : \overline{J} \times E \to \mathbb{R}$ solves*

$$\partial_t u(t,x) + (\mathcal{A}u)(t,x) - r(x)u(t,x) = 0, \tag{8.3.2}$$

for all $t \in J$ and $x \in E$, with boundary condition $u(T,x) = g(x)$, for all $x \in E$ and some function $g : E \to \mathbb{R}$, if and only if $u : \overline{J} \times E \to \mathbb{R}$ can be written as

$$u(t,x) = \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r(X(s))ds} g(X(T)) \middle| X(t) = x\right], \tag{8.3.3}$$

$\mathbb{Q}$-a.s., for all $t \in \overline{J}$ and $x \in E$.

Proof. Because of (8.3.2), the process M defined in (8.3.1) is by Proposition 8.13 a $\mathbb{Q}$-martingale such that

$$\begin{aligned} u(t,x) &= \mathbb{E}^{\mathbb{Q}}\left[M(t)|X(t) = x\right] = \mathbb{E}^{\mathbb{Q}}\left[M(T)|X(t) = x\right] \\ &= \mathbb{E}\left[e^{-\int_t^T r(X(s))ds} V(T, X(T)) \middle| X(t) = x\right] \\ &= \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r(X(s))ds} g(X(T)) \middle| X(t) = x\right], \end{aligned}$$

$\mathbb{Q}$-a.s., for all $t \in [0,T]$ and $x \in E$. For the other direction, we observe that every $u \in C^{1,2}(J \times E) \cap C^0(\overline{J} \times E)$ of the form (8.3.3) solves the PDE in (8.3.2) with boundary condition $u(T,x) = g(x)$, for all $x \in E$. □

Theorem 8.14 can only be used to describe the value function of an European option, but is not applicable for more general path-dependent portfolios such

as Asian, barrier or lookback options. In the following, we apply Theorem 8.14 to the one-dimensional Black-Scholes model with $d = m = 1$ and

$$dX(t) = rX(t)dt + \sigma X(t)dB(t),$$

for $t \in [0,T]$, where $r \in \mathbb{R}$ (constant interest rate) and $\sigma > 0$ (constant volatility), and where the state space is equal to $E =]0,\infty[$. In this case, the infinitesimal generator is given by $(\mathcal{A}f)(x) = rxf'(x) + \frac{1}{2}\sigma^2x^2f''(x)$, for all $f \in C^2(E)$. Using Theorem 8.14, we see that the discounted price of an European option with payoff function $g :]0,\infty[\to \mathbb{R}$ is a solution of the Black-Scholes PDE in (8.2.3) with terminal condition $u(T,x) = g(x)$, for all $x \in E$. In order to obtain a partial differential equation with non-degenerate operator and constant coefficients, we switch to log-prices such that Ito's formula in Theorem 1.39 implies

$$d\log X(t) = \left(r - \frac{1}{2}\sigma^2\right)dt + \sigma dB(t),$$

for $t \in [0,T]$, whose infinitesimal generator is then $(\mathcal{A}f)(x) = \left(r - \frac{1}{2}\sigma^2\right)f'(x) + \frac{1}{2}f''(x)$, for all $f \in C^2(E)$. In addition, we change to time-to-maturity $t \mapsto t-T$ to obtain a forward parabolic partial differential equation. Hence, by setting the function $v(t,x) = u(T-t,\log x)$, the Black-Scholes PDE in (8.2.3) becomes

$$\partial_t v(t,x) - \frac{1}{2}\sigma^2\partial_x^2 v(t,x) - \left(r - \frac{1}{2}\sigma^2\right)\partial_x v(t,x) + rv(t,x) = 0,$$

for all $(t,x) \in [0,T]\times]0,\infty[$, with initial condition $v(0,x) = g(\exp(x))$.

Altogether, the problem of pricing an European option written on an E-valued Ito diffusion $X = (X(t))_{t\in[0,T]}$ is equivalent to solving the corresponding PDE, which is also possible for local volatility models of the form

$$dX(t) = rX(t)dt + \xi(X(t))X(t)dB(t),$$

for $i = 1,...,d$ and $t \in [0,T]$, where $r \in \mathbb{R}$ (constant interest rate) and $\xi : E \to \mathbb{R}$ is a deterministic function of X. Changing with Ito's formula in Theorem 1.39 to log-prices, where the infinitesimal generator is $(\mathcal{A}f)(x) = \left(b - \frac{1}{2}\xi(\exp(x))^2\right)f'(x) + \frac{1}{2}\xi(\exp(x))^2f''(x)$, for all $f \in C^2(E)$, we obtain the parabolic PDE

$$\partial_t v(t,x) - \frac{1}{2}\xi(\exp(x))^2\partial_x^2 v(t,x) - \left(r - \frac{1}{2}\xi(\exp(x))^2\right)\partial_x v(t,x) + rv(t,x) = 0,$$

with boundary condition $v(0,x) = g(\exp(x))$, describing the discounted value function of the European option $g(X(T))$ in this local volatility model. For more details and in particular the PDE corresponding to other processes, we

refer to [HRSW13], where we emphasize that the PDEs considered in this section are all linear. Opposed to this, we consider in the next chapter a particular non-linear PDE, which can be used to price American options or European options in incomplete markets.

Chapter 9

Hamilton-Jacobi-Bellman Equation

In this chapter, we examine particular non-linear second order parabolic partial differential equations (PDEs) in more detail, where the Hamiltonian characterizes the evolution of the semigroup associated to the PDE. Subsequently, we focus on a special type of these non-linear second order parabolic PDE, which are of the Hamilton-Jacobi-Bellman (HJB) type and often appear in stochastic optimal control theory. After providing some main notions and results regarding the HJB equation such as verification theorems and viscosity solutions, we consider particular financial applications of the HJB equation. We follow the ideas of [GH14] to get the mathematical intuition and refer to [Pha09] and [Tou13] for rigorous proofs.

9.1 Hamiltonian and Comparison Principle

For an open subset $E \subset \mathbb{R}^d$, we consider throughout this chapter the second order fully non-linear PDE of the form

$$-\partial_t u(t,x) - H(t,x,u(t,x),\nabla u(t,x),\nabla^2 u(t,x)) = 0, \tag{9.1.1}$$

for all $(t,x) \in [0,T[\times E$, with terminal condition $u(T,x) = g(x)$, for all $x \in E$. Hereby, the function $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}_+^d \to \mathbb{R}$ with $(t,x,z,p,r) \mapsto H(t,x,z,p,r)$ is called the *Hamiltonian*, where $\mathbb{S}_+^d$ denotes the space of symmetric positive definite matrices. By inverting time $s = T - t$ and defining the function $v : [0,T] \times E \to \mathbb{R}$ by $v(s,x) = u(T-s,x)$, we obtain the initial value problem

$$-\partial_s v(s,x) - H(T-s,x,v(s,x),\nabla v(s,x),\nabla^2 v(s,x)) = 0,$$

for all $(s,x) \in]0,T] \times E$, with initial condition $v(0,x) = g(x)$, for all $x \in E$. As we have seen in Section 8.2, the PDE (9.1.1) is parabolic if $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}_+^d \to \mathbb{R}$ is elliptic, which means that $\left(\frac{\partial H}{\partial r_{ij}}\right)_{i,j=1,\dots,d}$ is symmetric positive semidefinite. Otherwise, if $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}_+^d \to \mathbb{R}$ is not differentiable, then $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}_+^d \to \mathbb{R}$ is parabolic if for every $t \in [0,T]$, $x \in E$, $z \in \mathbb{R}$, $p \in \mathbb{R}^d$, $r_1, r_2 \in \mathbb{S}_+^d$ with matrix $r_2 - r_1$ positive semidefinite, it follows that $H(t,x,z,p,r_2) \geqslant H(t,x,z,p,r_1)$.

Two important special cases of the non-linear PDE in (9.1.1) should be mentioned. If $H(t,x,z,p,r) = \sum_{i,j=1}^d a_{ij}(t,x) r_{ij} - f(t,x,z,p)$, for some functions $a : [0,T] \times E \to \mathbb{R}^{d \times d}$ and $f : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d$, then the PDE

$$-\partial_t u(t,x) - \sum_{i,j=1}^d a_{ij}(t,x) \partial_{ij} u(t,x) = f(t,x,u(t,x),\nabla u(t,x)), \tag{9.1.2}$$

is called *semilinear*, which means that $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}_+^d \to \mathbb{R}$ is linear in $\nabla^2 u$. Otherwise, if the functions a_{ij} in (9.1.2) additionally depend on u and ∇u, the PDE is called *quasilinear*.

In the following, we establish conditions such that the PDE in (9.1.1) satisfies the comparison principle. In general, a partial differential equation is said to satisfy the *comparison principle* if for two solutions $u, v : [0,T] \times E \to \mathbb{R}$ of the PDE to different terminal conditions $u(T,x) = g(x) \geqslant h(x) = v(T,x)$, for all $x \in E$, it follows that $u(t,x) \geqslant v(t,x)$, for all $(t,x) \in [0,T[\times E$. For example the value of an European option with payoff function $g : E \to \mathbb{R}$ dominating another European option with payoff function $h : E \to \mathbb{R}$ satisfies this principle, which is a consequence of the absence of arbitrage. From this, we conclude for two solutions $u, v : [0,T] \times E \to \mathbb{R}$ having the same terminal condition $u(T,x) = v(T,x) = g(x)$, for all $x \in E$, that $u(t,x) = v(t,x)$ coincides, for all $(t,x) \in [0,T] \times E$. Now, we find conditions on the Hamiltonian $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}_+^d \to \mathbb{R}$ such that the associated PDE in (9.1.1) satisfies the comparison principle.

Lemma 9.1. *The following holds true for a Hamiltonian* $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d_+ \to \mathbb{R}$ *given by* $[0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d_+ \ni (t,x,z,p,r) \mapsto H(t,x,z,p,r) \in \mathbb{R}$.

(i) If H *is parabolic and independent of* z*, then* (9.1.1) *satisfies the comparison principle.*

(ii) If H *is parabolic and strictly decreasing in* z*, then* (9.1.1) *satisfies the comparison principle.*

(iii) If H *is parabolic and there exists some* $C > 0$ *such that*

$$H(t,x,z_2,p,r) - H(t,x,z_1,p,r) \leqslant C(z_2 - z_1),$$

for all $t \in [0,T]$*,* $x \in E$*,* $z_1, z_2 \in \mathbb{R}$*,* $p \in \mathbb{R}^d$*,* $r \in \mathbb{S}^d_+$*, then* (9.1.1) *satisfies the comparison principle.*

Proof. This follows by [Tou13, Theorem 5.21], see also [GH14, Proposition 4.1] for a simpler proof when the state space $E \subset \mathbb{R}^d$ is compact. □

In a similar way, there exists a comparison principle for supersolutions and subsolutions of the PDE in (9.1.1). A function $u : [0,T] \times E \to \mathbb{R}$ is a *supersolution (resp. subsolution)* of the PDE (9.1.1) if

$$-\partial_t u(t,x) - H(t,x,u(t,x),\nabla u(t,x),\nabla^2 u(t,x)) \geqslant 0 \quad (\text{resp. } \leqslant 0),$$

for all $(t,x) \in [0,T] \times E$. By Ito's formula, it follows that if $t \mapsto u(t,X(t))$ is a (sub-) supermartingale with E-valued Ito diffusion $X = (X(t))_{t\in[0,T]}$ having infinitesimal generator $\mathcal{A}$, then $u : [0,T] \times E \to \mathbb{R}$ is a (sub-) supersolution of $-\partial_t u(t,x) - (\mathcal{A}u)(t,x) = 0$, which explains both minus signs in (9.1.1). Using supersolutions and subsolutions, a general PDE satisfies the *second comparison principle* if for every supersolution $u : [0,T] \times E \to \mathbb{R}$ and subsolution $v : [0,T] \times E \to \mathbb{R}$ with $u(T,x) = g(x) \geqslant h(x) = v(T,x)$, for all $x \in E$, it follows that $u(t,x) \geqslant v(t,x)$, for all $(t,x) \in [0,T] \times E$. Hereby, [Tou13, Theorem 5.21] shows that the PDE (9.1.1) satisfies the second comparison principle if one condition in Lemma 9.1 holds true.

In the following, we shortly outline the connection of the PDE in (9.1.1) to the pricing map of a financial derivative. As in [AGLM93] and [Bar94], we denote by $T_{t,t+h}$ the pricing functor assigning the fair price $u(t+h,\cdot)$ of a payoff at time $t+h$, with $h \in [-t,0]$ because we proceed from the terminal condition $u(T,\cdot) = g$ backwards in time. We collect the pricing functors in the semigroup $(T_{s,t})_{t\leqslant s}$ satisfying $T_{t,u}T_{u,s} = T_{t,s}$ and $T_{t,t} = \mathrm{id}$, and denote by $\mathcal{U}$ the set of value functions of European options, see [AGLM93] and [Bar94].

Axiom 9.2 (Causality Axiom). *There exists a semigroup $(T_{s,t})_{t\leqslant s}$ of operators $T_{s,t} : \mathcal{U} \to \mathcal{U}$ such that $u(t+h,x) = T_{t,t+h}(u(t,\cdot))(x)$, for all $t \in [0,T]$, $h \in [-t,0]$ and $u \in \mathcal{U}$. Then, it follows that*

$$\frac{u(t+h,x)-u(t,x)}{h} = \frac{T_{t,t+h}(u(t,\cdot))(x) - T_{t,t}(u(t,\cdot))(x)}{h},$$

which implies by letting $h \to 0^-$ that $\partial_t u(t,x) = H(u(t,\cdot))(x)$ such that the Hamiltonian $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d_+ \to \mathbb{R}$ describes the infinitesimal generator of the semigroup $(T_{t,s})_{s\leqslant t}$.

Axiom 9.3 (Monotonicity Axiom). *In order to exclude arbitrage opportunities, we impose that the semigroup $(T_{s,t})_{t\leqslant s}$ satisfies a comparison principle in the sense that $T_{t,t+h}(u)(t,x) \leqslant T_{t,t+h}(v)(t,x)$, for all $t \in [0,T]$, $h \in [-t,0]$, $x \in E$ and functions $u, v \in \mathcal{U}$ with $u(t,x) \leqslant v(t,x)$, for all $x \in E$.*

Axiom 9.4 (Locality Axiom). *If two functions $u, v \in \mathcal{U}$ are smooth and satisfy $\partial^\alpha u(t,x) = \partial^\alpha v(t,x)$, for all $\alpha \in \mathbb{N}^d$ and some $(t,x) \in [0,T[\times E$, then it follows that $T_{t,t+h}(u(t,\cdot))(x) - T_{t,t+h}(v(t,\cdot))(x) = o(h)$, as $h \to 0^-$. This condition ensures that the infinitesimal generator of the semigroup $(T_{s,t})_{t\leqslant s}$ given by the Hamiltonian $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d_+ \to \mathbb{R}$ is a local operator in the sense that*

$$\lim_{h\to 0^-} \frac{T_{t,t+h}(u(t,\cdot))(x) - u(t,x)}{h} = H(t,x,\nabla u(x), ..., \nabla^k u(x), ...),$$

which means that the infinitesimal generator of the semigroup $(T_{s,t})_{t\leqslant s}$ depends only on the derivatives of its argument $u : [0,T] \times E \to \mathbb{R}$ at the point x.

Axiom 9.5 (Regularity Axiom). *If $u, v \in \mathcal{U}$ are smooth, then for every $t \in [0,T]$ there exists some $C > 0$ such that*

$$\sup_{x\in E} |T_{t,t+h}(u(t,\cdot) + \lambda v(t,\cdot))(x) - T_{t,t+h}(u(t,\cdot))(x) - \lambda T_{t,t+h}(v(t,\cdot))(x)| \leqslant C\lambda h,$$

for all $t \in [0,T]$, $h \in [-t,0]$ and $\lambda \in \mathbb{R}$. This condition ensures that the infinitesimal generator of the semigroup $(T_{s,t})_{t\leqslant s}$ given by $H : [0,T] \times E \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d_+ \to \mathbb{R}$ depends only on u, ∇u and $\nabla^2 u$.

Axiom 9.6 (Cash-Invariance Axiom)**.** *For all $u \in \mathcal{U}$ and $t \in [0,T]$, $h \in [-t,0]$ and $c \in \mathbb{R}$, it holds that*

$$T_{t,t+h}(u(t,\cdot)+c) = T_{t,t+h}(u(t,\cdot))+c, \quad T_{t,t+h}(u(t,\cdot+c)) = T_{t,t+h}(u(t,\cdot))(\cdot+c),$$

where the first (resp. second) condition implies that the Hamiltonian $H : [0,T]\times E\times\mathbb{R}\times\mathbb{R}^d\times\mathbb{S}^d_+ \to \mathbb{R}$ is independent of u (resp. x). However, the first condition is not justified by any financial interpretation.

From the nature of the conditions in Axioms 9.2-9.6, we expect that $(T_{s,t})_{t\leqslant s}$ describes the semigroup of the parabolic PDE in (9.1.1) with Hamiltonian $H : [0,T]\times E\times\mathbb{R}\times\mathbb{R}^d\times\mathbb{S}^d_+ \to \mathbb{R}$.

Theorem 9.7 ([AGLM93, Theorem 2])**.** *Under the Axioms 9.2-9.5, there exists a continuous function $H : [0,T]\times E\times\mathbb{R}\times\mathbb{R}^d\times\mathbb{S}^d_+ \to \mathbb{R}$ such that $u : [0,T] \to \mathbb{R}$ is a (viscosity) solution of*

$$-\partial_t u(t,x) - H(t,x,u,\nabla u,\nabla^2 u) = 0,$$

for all $(t,x) \in [0,T[\times E$, where $H : [0,T]\times E\times\mathbb{R}\times\mathbb{R}^d\times\mathbb{S}^d_+ \to \mathbb{R}$ satisfies the assumption of Lemma 9.1 such that (9.1.1) *fulfills the comparison principle. If we additionally assume Axiom 9.6, then $H : [0,T]\times\mathbb{R}^d\times\mathbb{S}^d_+ \to \mathbb{R}$ is independent of u and x such that*

$$-\partial_t u(t,x) - H(t,\nabla u,\nabla^2 u) = 0,$$

for all $(t,x) \in [0,T[\times E$, and $H : [0,T]\times\mathbb{R}^d\times\mathbb{S}^d_+ \to \mathbb{R}$ is parabolic.

In addition, we observe that the absence of arbitrage implies that the value function $u : E \to \mathbb{R}$ with terminal condition $u(T,x) = x_i$, i.e. the value of the discounted stock price (where we assume zero interest rates for simplicity), must satisfy $u(t,x) = x_i$ such that $H(t,e_i,0) = 0$, for all $t \in [0,T]$ and $i = 1,...,d$, where e_i denotes the ith unit vector of $\mathbb{R}^d$. As mentioned in Section 8.2, second order parabolic PDEs can be solved numerically by the finite difference or finite element method, see [LT05].

9.2 Stochastic Optimal Control

In the previous section, we considered parabolic second-order PDEs of the form (9.1.1), where we now focus on a special type of this class of PDEs, the so-called

Hamilton-Jacobi-Bellman (HJB) equation of the form

$$-\partial_t u(t,x) - \sup_{c\in\mathcal{U}}(\mathcal{A}^c u)(t,x) = 0, \tag{9.2.1}$$

for all $(t,x) \in [0,T] \times E$ with terminal condition $u(T,x) = g(x)$, for all $x \in E$, where $\mathcal{A}^c$ is described below. The HJB equation often appears in stochastic optimal control problems, for which we let $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{Q})$ be a filtered probability space with filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,\infty[}$ satisfying the usual condition and assume that $B = (B(t))_{t\in[0,T]}$ is a d-dimensional Brownian motion defined on $(\Omega, \mathcal{A}, \mathbb{F}, \mathbb{Q})$. For some $U \subset \mathbb{R}^n$, we denote by $\mathcal{U}$ the space of all $\mathbb{F}$-progressively measurable and U-valued processes $c = (c(t))_{t\in[0,T]}$, which will serve as control processes. For every control process $c \in \mathcal{U}_0$, let $X^c = (X^c(t))_{t\in[0,T]}$ be an E-valued solution of

$$dX^c(t) = b(t, X^c(t), c(t))dt + \xi(t, X^c(t), c(t))dB(t), \tag{9.2.2}$$

for $t \in [0,T]$ and some initial value $x_0 \in E$, where $E \subset \mathbb{R}^d$, for which we assume that the drift and volatility function $b : [0,T] \times E \times U \to \mathbb{R}^d$ and $\xi : [0,T] \times E \times U \to \mathbb{R}^{d\times d}$ satisfy a uniform Lipschitz and linear growth condition, i.e. there exists a constant $C > 0$ such that for every $t \in [0,T]$, $x, y \in E$ and $c \in U$, it holds that

$$\begin{aligned} \|b(t,x,c) - b(t,y,c)\| + \|\xi(t,x,c) - \xi(t,y,c)\| &\leqslant C\|x-y\|, \\ \|b(t,x,c)\| + \|\xi(t,x,c)\| &\leqslant C(1+\|x\|). \end{aligned} \tag{9.2.3}$$

In this case, $X^c = (X^c(t))_{t\in[0,T]}$ is called the *controlled process*, which is governed by the *control process* $c = (c(t))_{t\in[0,T]}$. Equation 9.2.3 ensures in particular that the SDE (9.2.2) admits an E-valued strong solution $X^c = (X^c(t))_{t\in[0,T]}$ such that $\mathbb{E}^{\mathbb{Q}}\left[\sup_{0\leqslant t\leqslant T}\|X^c(t)\|^2\right] < \infty$, for all $c \in \mathcal{U}$. This is a more general existence result than Theorem 8.1, for which we refer to [Pro05] for a proof with random coefficients.

For some $f, k : [0,T] \times E \times U \to \mathbb{R}$ and $g : E \to \mathbb{R}$, we define the cost functional $J : [0,T] \times E \times \mathcal{U} \to \mathbb{R}$ by

$$\begin{aligned} J(t,x,c) = \mathbb{E}^{\mathbb{Q}}\Bigg[\int_t^T e^{-\int_t^s k(r,X^c(r),c(r))dr} f(s, X^c(s), c(s))ds + \dots \\ \dots + e^{-\int_t^T k(r,X^c(r),c(r))dr} g(X^c(T)) \Bigg| X^c(t) = x \Bigg], \end{aligned} \tag{9.2.4}$$

for $t \in [0,T]$, $x \in E$ and $c \in \mathcal{U}$, where we assume that $k_-(\cdot) := \max(-k(\cdot), 0)$ is uniformly bounded in x. Moreover, we define $\mathcal{U}_{t,T}$ as the space of all $\mathbb{F}$-

progressively measurable processes $(c(s))_{t\leqslant s\leqslant T}$ such that $c(s)\in U$, for all $s\in[t,T]$, and $\mathbb{E}^{\mathbb{Q}}\left[\int_0^T|f(t,X^c(t),c(t))|dt+|g(X^c(T))|\right]<\infty$, which ensures that the cost functional is finite. Then, the corresponding stochastic optimal control problem consists of maximizing the cost functional $J:[0,T]\times E\times\mathcal{U}\to\mathbb{R}$, from which we define the value function $u:[0,T]\times E\to\mathbb{R}$ by

$$u(t,x)=\operatorname*{ess\,sup}_{c\in\mathcal{U}_{t,T}}J(t,x,c), \tag{9.2.5}$$

for $(t,x)\in[0,T[\times E$. In the following, we show the Bellman principle encoding the dynamic programming principle, where we use the standard form of (9.2.4) obtained by augmenting the state variable with the controlled processes $Y^c=(Y^c(t))_{t\in[0,T]}$ and $Z^c=(Z^c(t))_{t\in[0,T]}$ given by $dY^c(t)=Z^c(t)f(t,X^c(t),c(t))$ and $dZ^c(t)=-Z^c(t)k(t,X^c(t),c(t))dt$ such that the value function in (9.2.5) satisfies $u(t,x)=\widetilde{u}(t,x,0,1)$, where

$$\widetilde{u}(t,x,y,z)=\operatorname*{ess\,sup}_{c\in\widetilde{\mathcal{U}}_{t,T}}\mathbb{E}^{\mathbb{Q}}\left[\widetilde{g}(\widetilde{X}^c(T))\Big|\widetilde{X}^c(t)=(x,y,z)\right],$$

for $t\in[0,T]$, where $\widetilde{x}=(x,y,z)$, $\widetilde{X}^c=(X^c,Y^c,Z^c)$ and $\widetilde{g}(x,y,z)=y+g(x)z$. Hence, every stochastic optimal control problem in (9.2.5) can be reduced to the standard form with $f(t,x,u)=k(t,x,u)=0$.

Theorem 9.8 ([GH14, Section 4.4.3])**.** *For every $t\in[0,T[$, $x\in E$ and stopping time τ with values in $[t,T]$, it follows that*

$$u(t,x)=\operatorname*{ess\,sup}_{c\in\mathcal{U}_{t,T}}\mathbb{E}^{\mathbb{Q}}\left[u(\tau,X^c(\tau))\big|X^c(t)=x\right],\quad\mathbb{Q}\text{-a.s.} \tag{9.2.6}$$

Proof. Without loss of generality, we show (9.2.7) only for stochastic optimal control problems in standard form and for both inequalities "$\leqslant$" and "$\geqslant$", where $\tau\in\mathcal{T}_{t,T}:=\{\tau\text{ stopp. time; }\tau\in[t,T],\ \mathbb{Q}\text{-a.s.}\}$ is assumed to be a stopping time. For "$\leqslant$", it follows by applying the tower property that

$$u(t,x)=\operatorname*{ess\,sup}_{c\in\mathcal{U}_{t,T}}\mathbb{E}^{\mathbb{Q}}\left[\mathbb{E}^{\mathbb{Q}}\left[g(X^c(T))|\mathcal{F}_\tau\right]\big|X^c(t)=x\right],\quad\mathbb{Q}\text{-a.s.}$$

By definition, we have $\mathbb{E}^{\mathbb{Q}}\left[g(X^c(T))|\mathcal{F}_\tau\right]\leqslant u(\tau,X^c(\tau))$, $\mathbb{Q}$-a.s., for all $c\in\mathcal{U}_{\tau,T}$, such that

$$u(t,x)\leqslant\operatorname*{ess\,sup}_{c\in\mathcal{U}_{t,T}}\mathbb{E}^{\mathbb{Q}}\left[u(\tau,X^c(\tau))|X^c(t)=x\right],\quad\mathbb{Q}\text{-a.s.}$$

For the reverse inequality "$\geqslant$", let $c\in\mathcal{U}_{t,T}$ be an optimal control process

for $u(t,x)$, fix some $\varepsilon > 0$ and assume that $(c_\varepsilon(s))_{\tau \leqslant s \leqslant T}$ is a control process between τ and T safisfying

$$u(\tau, X^c(\tau)) \leqslant \mathbb{E}^{\mathbb{Q}}\left[g(X^{c_\varepsilon}(T))\middle|\mathcal{F}_\tau\right] + \varepsilon, \quad \mathbb{Q}\text{-a.s.} \tag{9.2.7}$$

From this, we define the glued control process[1] $\widetilde{c}_\varepsilon \in \mathcal{U}_{t,T}$ given by $\widetilde{c}_\varepsilon(s) = \mathbb{1}_{[t,\tau[}c(s) + \mathbb{1}_{[\tau,T]}c_\varepsilon(s)$, for $s \in [t,T]$. Applying the tower property and using (9.2.7), we conclude that

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left[u(\tau, X^c(\tau))\middle|X^c(t) = x\right] &\leqslant \mathbb{E}^{\mathbb{Q}}\left[\mathbb{E}^{\mathbb{Q}}\left[g(X^{c_\varepsilon}(T))\middle|\mathcal{F}_\tau\right]\middle|X^c(t) = x\right] + \varepsilon \\
&= \mathbb{E}^{\mathbb{Q}}\left[\mathbb{E}^{\mathbb{Q}}\left[g(X^{\widetilde{c}}(T))\middle|\mathcal{F}_\tau\right]\middle|X^c(t) = x\right] + \varepsilon \\
&= \mathbb{E}^{\mathbb{Q}}\left[g(X^{\widetilde{c}}(T))\middle|X^c(t) = x\right] + \varepsilon \\
&\leqslant u(t,x) + \varepsilon, \quad \mathbb{Q}\text{-a.s.}
\end{aligned}$$

Letting $\varepsilon \to 0$ shows the reverse inequality "$\geqslant$" such that (9.2.6) follows. □

Theorem 9.8 is known as the dynamic programming equation, which means that if a control process is optimal for the final value function, it must be optimal at each intermediate time. In the following, we show that the value function of an stochastic optimal control problem solves the HJB equation.

Theorem 9.9 ([Tou13, Theorem 6.4])**.** *The optimal value function $u : [0,T] \times E \to \mathbb{R}$ in (9.2.5) is a (viscosity) solution of*

$$-\partial_t u(t,x) - \sup_{c \in U}\left((\mathcal{A}^c u)(t,x) + f(t,x,c) - k(t,x,c)u(t,x)\right) = 0, \tag{9.2.8}$$

for all $(t,x) \in [0,T[\times E$ with terminal condition $u(T,x) = g(x)$, for all $x \in E$, where $\mathcal{A}^c = (\mathcal{A}^c_t)_{t\in[0,T]}$ is the infinitesimal generator of (9.2.2) given by $\mathcal{A}^c_t = \sum_{i=1}^d b_i(t,x,c)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^d \sigma_{ij}(t,x,c)\frac{\partial^2}{\partial x_i \partial x_j}$, where $\sigma(t,x,c) = \xi(t,x,c)\xi(t,x,c)^\top$, for $(t,x,c) \in [0,T] \times E \times U$.

For more details regarding verification results, we refer to [Tou13] and [Pha09], whereas [Eva15] and [Kry87] gives existence and uniqueness results for the HJB equation in (9.2.8) together with their regularity.

[1] This needs a technical argument involving a measurability selection arguments, see [Tou13, Section 2.2.2+2.3].

9.3 Pricing Equation of American Option

In this section, we consider American options giving the owner the right to exercise a financial derivative at any time. These options form a typical non-linear problem, whose value function is a solution of the Hamilton-Jacobi-Bellman (HJB) equation. From the basic pricing equation, we consider the optimal stopping problem and show the variational inequality of the American option. We assume that the stock price process follows an Ito diffusion X with infinitesimal generator $\mathcal{A}$ under an equivalent martingale measure $\mathbb{Q}$, where X is a solution of the SDE (8.1.1), whose coefficients satisfy the assumptions of Theorem 8.1.

Let $\mathcal{T}_{t,T}$ be the set of stopping times τ such that $\tau \in [t,T]$, $\mathbb{Q}$-a.s. Then, the owner of the American option with payoff function $g : E \to \mathbb{R}$ tries to maximize the payoff of the option by exercising the contract at best time, i.e.

$$u(t,x) = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^{\mathbb{Q}} \left[e^{-\int_t^\tau r(s,X(s))ds} g(X(\tau)) \Big| X(t) = x \right], \quad \mathbb{Q}\text{-a.s.}, \tag{9.3.1}$$

for $t \in [0,T]$ and $x \in E$, where $r : [0,T] \times E \to \mathbb{R}$ denotes the short term interest rate, and where it holds that $u(t,x) \geqslant g(x)$, for all $(t,x) \in [0,T] \times E$. More precisely, for strict inequality $u(t,x) > g(x)$, the holder shoud continue to hold the option (the so-called *continuation region*) and as soon as equality arises $u(\tau,x) = g(x)$, where $\tau \in \mathcal{T}_{t,T}$ denotes a stopping time, the owner should exercise the option (the so-called *exercise region*). Hereby, we observe that the optimal stopping time τ is given by

$$\tau^* = \inf \{ s \geqslant t;\, u(s, X(s)) = g(X(s)) \}. \tag{9.3.2}$$

In the following, we state the variational inequality corresponding to the value function of an American option in (9.3.1), where we refer to [BL82] and [JLL90] for more details.

Theorem 9.10 ([GH14, Section 4.5]). *Assume that $v \in C^{1,2}([0,T[\times E) \cap C^0([0,T] \times E)$ solves the variational inequality*

$$\max \left(\partial_t v(t,x) + (\mathcal{A}v)(t,x) - r(t,x)v(t,x), g(x) - v(t,x) \right) = 0, \tag{9.3.3}$$

for all $(t,x) \in [0,T[\times E$, with terminal condition $v(T,x) = g(x)$, for all $x \in E$, then $v(t,x) = u(t,x)$, for all $(t,x) \in [0,T] \times E$, where $u : [0,T] \times E \to \mathbb{R}$ is given by (9.3.1).

Proof. First, Proposition 8.13 shows that the process $M = (M(s))_{s \in [t,T]}$ defined by $M(s) = v(x, X(s)) - \int_t^s (\partial_t - \mathcal{A} - r)\, v(u, X(u)) du$ is a true martin-

gale such that the optional stopping theorem implies for a stopping time $\tau \in \mathcal{T}_{t,T}$ that $\mathbb{E}^{\mathbb{Q}}[M(\tau)|\mathcal{F}_t] = M(t) = v(t, X(t))$, $\mathbb{Q}$-a.s. Since $\partial_t v(s, X(s)) + (\mathcal{A}v)(s, X(s)) - r(s, X(s))v(s, X(s)) \leqslant 0$ and $v(\tau, X(\tau)) \geqslant g(X(\tau))$, $\mathbb{Q}$-a.s., by (9.3.3), we have $v(t, X(t)) \geqslant \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^\tau r(s,X(s))ds} g(X(\tau))\middle|\mathcal{F}_t\right]$, $\mathbb{Q}$-a.s., for all $\tau \in \mathcal{T}_{t,T}$, and therefore

$$v(t, X(t)) \geqslant \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^\tau r(s,X(s))ds} g(X(\tau))\middle|\mathcal{F}_t\right], \quad \mathbb{Q}\text{-a.s.}, \tag{9.3.4}$$

for all $t \in [0, T]$. On the other hand, applying the optional stopping theorem to the stopping time τ^* in (9.3.2), it follows that $\mathbb{E}^{\mathbb{Q}}[M(\tau^*)|\mathcal{F}_t] = M(t) = u(t, X(t))$, $\mathbb{Q}$-a.s. Since $\partial_t v(s, X(s)) + (\mathcal{A}v)(s, X(s)) - r(s, X(s))v(s, X(s)) = 0$, for all $s \in [t, \tau^*]$ and $v(\tau^*, X(\tau^*)) = g(X(\tau^*))$, $\mathbb{Q}$-a.s., it follows that

$$v(t, X(t)) = \mathbb{E}^{\mathbb{Q}}[g(X(\tau^*))|\mathcal{F}_t], \quad \mathbb{Q}\text{-a.s.}$$

With (9.3.4), we have $v(t, X(t)) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^\tau r(s,X(s))ds} g(X(\tau))\middle|\mathcal{F}_t\right]$, $\mathbb{Q}$-a.s., for all $t \in [0, T]$, where the essential supremum over $\mathcal{T}_{t,T}$ is attained for the optimal stopping time τ^* in (9.3.2). □

9.4 Super-Replication Price in Stochastic Volatility Model

In this section, we consider the problem of pricing and hedging options in an incomplete market modelled by the Jacobi stochastic volatility model, see [AFP18]. We assume that the interest rate is constant and the only risky asset $X = (X(t))_{t\in[0,T]}$ is under the real-world measure $\mathbb{P}$ a solution of the SDE

$$\begin{aligned} dX(t) &= rX(t)dt + \rho X(t)\sqrt{Q(\sigma(t))}dB_1(t) + X(t)\sqrt{\sigma(t) - \rho^2 Q(\sigma(t))}dB_2(t), \\ d\sigma(t) &= \kappa(\theta - \sigma(t))dt + \xi\sqrt{Q(\sigma(t))}dB_1(t), \end{aligned} \tag{9.4.1}$$

for $t \in [0, T]$, with initial values $X(0) > 0$ and $\sigma(0) \in [a, b]$, where $0 \leqslant a \leqslant b < \infty$, where B_1 and B_2 are independent Brownian motions and $\sigma = (\sigma(t))_{t\in[0,T]}$ denotes the stochastic volatility. In addition, the parameters satisfy $r \in \mathbb{R}$, $\kappa > 0$, $\theta \in]a, b]$, $\xi > 0$ and $\rho \in [-1, 1]$, and the function $Q : [a, b] \to \mathbb{R}$ is given by

$$Q(v) = \frac{(b - v)(v - a)}{(\sqrt{b} - \sqrt{a})^2},$$

for $v \in [a,b]$. Then, it follows by [AFP18] that (9.4.1) admits a unique solution taking values in $]0,\infty[\times[a,b]$ such that no state $v \in [a,b[$ is absorbing. Moreover, σ takes values in $]a,b[$ if and only if $\sigma(0) \in]a,b[$ and

$$\xi^2 \frac{b-a}{(\sqrt{b}-\sqrt{a})^2} \leqslant 2\kappa \min(b-\theta, \theta-a). \tag{9.4.2}$$

For $\sigma(0) = b = \theta$, we obtain the Black-Scholes model and for $a \to 0$ and $b \to \infty$, we recover the Heston model in [Hes93], where (9.4.2) becomes the Feller condition $\xi^2 \leqslant 2\kappa\theta$. Moreover, the infinitesimal generator is given by

$$\begin{aligned}(\mathcal{A}f)(x,y) = \frac{1}{2}x^2 y \partial_{xx} f(x,y) + \rho\xi x Q(y) \partial_{xy} f(x,y) + \frac{1}{2}\xi^2 Q(y) \partial_{yy} f(x,y) \\ + rx\partial_x f(x,y) + \kappa(\theta - y)\partial_y f(x,y),\end{aligned} \tag{9.4.3}$$

for $f \in C^2(]0,\infty[\times[a,b])$ and $(x,y) \in]0,\infty[\times[a,b]$. In addition, we observe that the discounted price process $\widetilde{X} = (D_{0t}X(t))_{t\in[0,T]}$ is a martingale under $\mathbb{P}$, although $\mathbb{P}$ is not the only martingale measure for (9.4.1).

In the following, we will always assume that $a > 0$ and $\rho^2 < 1$ as well as that the conditions in [AFP18, Theorem 3.1] are satisfied. For some $\lambda \in \mathbb{R}$, we define the two-dimensional process $\eta = (\eta(t))_{t\in[0,T]}$ by

$$\eta_1(t) = \frac{\lambda}{\xi\sqrt{Q(\sigma(t))}} \qquad \text{and} \qquad \eta_2(t) = -\frac{\lambda\rho}{\xi\sqrt{\sigma(t) - \rho^2 Q(\sigma(t))}},$$

for $t \in [0,T]$. Using this, we define the equivalent martingale measure $\mathbb{Q}_\lambda \sim \mathbb{P}$ via its density process by

$$\mathbb{E}\left[\frac{d\mathbb{Q}_\lambda}{d\mathbb{P}}\middle|\mathcal{F}_t\right] = \exp\left(-\int_0^t \eta_1(s)dB_1(s) - \int_0^t \eta_2(s)dB_2(s) - \frac{1}{2}\int_0^t \|\eta(s)\|^2 ds\right),$$

for $t \in [0,T]$, such that the processes $B_1^{\mathbb{Q}_\lambda}$ and $B_2^{\mathbb{Q}_\lambda}$ defined by $dB_1^{\mathbb{Q}_\lambda}(t) = dB_1(t) + \eta_1(t)dt$ and $dB_2^{\mathbb{Q}_\lambda}(t) = dB_2(t) + \eta_2(t)dt$ are by Girsanov's theorem in [KS98, Theorem 3.5.1] independent Brownian motions under $\mathbb{Q}_\lambda$. Inserting this into (9.4.1), we observe that X and σ follow under $\mathbb{Q}_\lambda$ the SDE

$$\begin{aligned}dX(t) &= rX(t)dt + \rho\sqrt{Q(\sigma(t))}dB_1(t) + \sqrt{\sigma(t) - \rho^2 Q(\sigma(t))}dB_2(t), \\ d\sigma(t) &= \kappa(\theta_\lambda - \sigma(t))dt + \xi\sqrt{Q(\sigma(t))}dB_1(t),\end{aligned}$$

for all $t \in [0,T]$, where $\theta_\lambda = \theta - \lambda/\kappa$. Because σ also takes values in $[a,b]$ under $\mathbb{Q}_\lambda$, we need that $\theta_\lambda \in]a,b]$ such that every $\lambda \in \Lambda := [\kappa(\theta-b), \kappa(\theta-a)[$ yields an equivalent martingale measure $\mathbb{Q}_\lambda \sim \mathbb{P}$.

In this case, the super-replication price (seller's price) of an European option with payoff function $g : E \to \mathbb{R}$ can be written as stochastic control problem, whose value function is given by

$$u_s(t,x,y) = \operatorname*{ess\,sup}_{\lambda\in\Lambda} \mathbb{E}^{\mathbb{Q}_\lambda}\left[e^{-r(T-t)}g(X(T))\Big|X(t)=x,\sigma(t)=y\right], \quad \mathbb{P}\text{-a.s.}, \tag{9.4.4}$$

for $(t,x,y) \in [0,T]\times]0,\infty[\times[a,b]$. Similarly, the sub-replication price (buyer's price) is given by

$$u_b(t,x,y) = \operatorname*{ess\,inf}_{\lambda\in\Lambda} \mathbb{E}^{\mathbb{Q}_\lambda}\left[e^{-r(T-t)}g(X(T))\Big|X(t)=x,\sigma(t)=y\right], \quad \mathbb{P}\text{-a.s.}, \tag{9.4.5}$$

for $(t,x,y) \in [0,T]\times]0,\infty[\times[a,b]$, which can be reformulated as standard stochastic optimal control problem (maximization problem) by using minus signs. Hence, the optimal value function $u_s : [0,T]\times]0,\infty[\times[a,b] \to \mathbb{R}$ in (9.4.4) satisfies by Theorem 9.9 the Hamilton-Jacobi-Bellman (HJB) equation

$$-\partial_t u_s(t,x,y) - \sup_{\lambda\in\Lambda}\left(\left(\mathcal{A}^\lambda u_s\right)(t,x,y) - r(t,x)u_s(t,x,y)\right) = 0, \tag{9.4.6}$$

for all $(t,x,y) \in [0,T[\times]0,\infty[\times[a,b]$, with terminal condition $u(T,x) = g(x)$, for all $x \in]0,\infty[$, where $\mathcal{A}^\lambda$ denotes the infinitesimal generator of $(X,\sigma) = (X(t),\sigma(t))_{t\in[0,T]}$ under $\mathbb{Q}_\lambda \sim \mathbb{P}$ given by

$$\begin{aligned}\left(\mathcal{A}^\lambda f\right)(x,y) = {}& \frac{1}{2}x^2 y\partial_{xx}f(x,y) + \rho\xi x Q(y)\partial_{xy}f(x,y) + \frac{1}{2}\xi^2 Q(y)\partial_{yy}f(x,y)\\ &+ rx\partial_x f(x,y) + (\kappa\theta - \lambda - \kappa y)\partial_y f(x,y),\end{aligned}$$

for $f \in C^2([0,\infty[\times[a,b])$ and $(x,y) \in]0,\infty[\times[a,b]$. On the other hand, we conclude with sign changes and Theorem 9.9 that the optimal value function $u_b : [0,T]\times]0,\infty[\times[a,b] \to \mathbb{R}$ in the sub-replication problem (9.4.5) satisfies the HJB equation

$$-\partial_t u_b(t,x,y) - \inf_{\lambda\in\Lambda}\left(\left(\mathcal{A}^\lambda u_b\right)(t,x,y) - r(t,x)u_b(t,x,y)\right) = 0, \tag{9.4.7}$$

for all $(t,x,y) \in [0,T[\times]0,\infty[\times[a,b]$, with terminal condition $u(T,x) = g(x)$, for all $x \in]0,\infty[$. By solving the HJB equations in (9.4.6) and (9.4.7), we obtain the super-replication and sub-replication price of an European option in this Jacobi stochastic volatility model, where the Jacobi model is a well tractable and versatile stochastic volatility model including the Black-Scholes model and Heston model as limit case.

Chapter 10

Abstract Neural Network and Numerical Examples

In this chapter, we use abstract neural networks between infinite dimensional spaces to learn the non-linear pricing operator of a financial derivative, which either has an American exercise style or is written on an asset price process following a stochastic volatility model. More precisely, we approximate the solution operator $T_{T,t}$ of the semigroup $(T_{s,t})_{t\leqslant s}$ associated to the Hamilton-Jacobi-Bellman (HJB) equation, which propagates the payoff function $g : E \to \mathbb{R}$ of the financial option (appearing in the terminal condition of the HJB equation) backwards to the value function $u(t, x)$ of the option at time $t < T$, corresponding to the price of the option at time $t < T$ if the price vector was $X(t) = x$. By showing that the solution operator $g \mapsto (t \mapsto u(t, \cdot) := T_{t,T}(g))$ is continuous between two function spaces, we use the universal approximation result for abstract neural networks in Theorem 3.27 to approximate the pricing operator.

In general, neural networks have been succesfully applied for pricing and hedging financial derivatives, see [RW19] for a literature review. In [EHJ17], neural networks are for example used to learn the solution of a parabolic PDE, whereas [GHJW18] showed explicitely for the Black-Scholes PDE that this method overcomes the curse of dimensionality. Opposed to these contributions, we apply abstract neural networks to learn the solution operator of a given PDE instead of only one solution. This means that we obtain the solution $u : [0, T] \times E \to \mathbb{R}$ by inserting the terminal condition (payoff function) $g : E \to \mathbb{R}$.

Moreover, we will see in the construction of the abstract neural network approximating the solution operator that the option prices of either the Black-Scholes model or the Bachelier model are universal in the sense that they can be used to learn the price of the same option, but which is either of American style or written on assets in an incomplete market given by a stochastic

volatility model. The method can be applied under some assumptions on the parameters to very general financial market models driven by an Ito diffusion as in Chapter 8.

10.1 Approximation of the Solution Operator

In order to simplify notation, we assume that the market consists of only one asset, which is modelled by a stochastic process with state space $E =]0, \infty[$, e.g. geometric Brownian motion or the Heston model. However, the following results also hold true for market models with multiple assets, where the state space is for example $E =]0, \infty[^d$. In the following, we introduce the function spaces, which are needed to construct the abstract neural network approximating the solution operator of either the variational inequality in (9.3.3) describing American option prices or the HJB equation in (9.4.6) (resp. (9.4.7)) representing the super-replication (resp. sub-replication) price of an European option in an incomplete market.

For the input space X containing the payoff functions, we define the weight $\gamma : E \to]0, \infty[$ by $\gamma(x) = \exp(-|x|/2)$, for $x \in E$. Using this, we consider the Hilbert space $L^2_\gamma(E) = \left\{ g \in L^2(E);\, \|g\gamma\|_{L^2(E)} < \infty \right\}$ with inner product $\langle f, g \rangle_{L^2_\gamma(E)} = \int_E f(x) g(x) \gamma(x)^2 dx$, for $f, g \in L^2_\gamma(E)$. Then, we observe that the map $J : L^2_\gamma(E) \to L^2_\gamma(E)^*$ defined by

$$J(g) = \left(f \mapsto \int_E \gamma(x)^2 f(x) g(x) dx \right),$$

for $g \in L^2_\gamma(E)$, is an isometric isomorphism such that $L^2_\gamma(E)^* \cong L^2_\gamma(E)$. Then, we define the input space $X := L^2_\gamma(E) \cong L^2_\gamma(E)^*$ equipped with the weak-$*$-topology and introduce the weight function $\psi : X \to]0, \infty[$ defined by

$$\psi(g) = \exp\left(\|g\|^2_{L^2_\gamma(E)} \right),$$

for $g \in X$, which is an admissible weight function, because every preimage $K_R := \psi^{-1}(]0, R])$ is by Banach-Alaoglu in Theorem 1.15 compact in the weak-$*$-topology, for all $R > 0$, such that (X, ψ) is a weighted space. In particular, all bounded functions $g : E \to \mathbb{R}$ and all piecewise polynomial payoff functions $g : E \to \mathbb{R}$ including the payoff functions of call and put options are contained in $L^2_\gamma(E)$.

For the output space Y, we consider the Hilbert space $L^2_\gamma([0, T] \times E \times F)$, where $F \subset \mathbb{R}$ denotes the compact subet of additional parameters such as the

stochastic volatility, which we equip with the inner product

$$\langle u, v\rangle_{L^2_\gamma([0,T]\times E\times F)} = \int_0^T \int_E \int_F \gamma(x)^2 u(t,x,y)v(t,x,y)dydxdt,$$

for $u, v \in Y$. In the case when there are no additional parameters for the value function, we omit the integration over F. On the other hand, we choose $Z = \mathbb{R}$ for the hidden layer space with weight function $\psi_Z : \mathbb{R} \to]0,\infty[$ defined by $\psi_Z(z) = \exp(|z|)$, for $z \in \mathbb{R}$, where every preimage $K_{Z,R} := \psi_Z^{-1}(]0,R])$ is as in Example 3.2(i) compact in $\mathbb{R}$.

In this setting, we consider either the variational inequality in (9.3.3) (for American options) or the HJB equation in (9.4.6) or (9.4.7) (for the super-/sub-replication problem of an European option). In the first case, we define the solution operator $\mathcal{S}_a : X \to Y$ as the function $u := \mathcal{S}_a(g)$, for $g \in L^2_\gamma(E)$, which satisfies $u(t,\cdot) \in \mathcal{K}_g := \{v \in L^2_\gamma(E);\, v \geqslant g \text{ a.e.}\}$, a.e., and solves the variational inequality

$$-\langle \partial_t u(t,\cdot) + \mathcal{A}u(t,\cdot) - ru(t,\cdot), v - u\rangle_{L^2_\gamma(E)} \leqslant 0, \tag{10.1.1}$$

for a.e. $t \in [0,T]$ and $v \in \mathcal{K}_g$, with terminal condition $u(T,x) = g(x)$, for all $x \in E$, where we assume that the interest rate $r \in \mathbb{R}$ is constant. For further details, we refer to [HRSW13, Section 5.2] for American options and [HRSW13, Appendix B] for variational inequalities.

On the other hand, we define the solution operators $\mathcal{S}_s, \mathcal{S}_b : X \to Y$ for the HJB equations representing the super-/sub-replication problem as the functions $u_s := \mathcal{S}_s(g)$ and $u_b := \mathcal{S}_b(g)$, which satisfy

$$\begin{aligned} -\partial_t u_s(t,x,y) - \sup_{\lambda\in\Lambda} \left(\left(\mathcal{A}^\lambda u_s\right)(t,x,y) - r(t,x)u_s(t,x,y)\right) &= 0, \\ -\partial_t u_b(t,x,y) - \inf_{\lambda\in\Lambda} \left(\left(\mathcal{A}^\lambda u_b\right)(t,x,y) - r(t,x)u_b(t,x,y)\right) &= 0, \end{aligned} \tag{10.1.2}$$

for all $(t,x,y) \in [0,T]\times]0,\infty[\times[a,b]$, with terminal condition $u_b(T,x,y) = u_s(T,x,y) = g(x)$, for all $(x,y) \in]0,\infty[\times[a,b]$, where we refer to [Tou13] and [Pha09] for further details. In the following, we show that all three solution operators $\mathcal{S}_a, \mathcal{S}_s, \mathcal{S}_b : X \to Y$ belong to the weighted function space $\mathcal{B}_\psi(X;Y)$.

Lemma 10.1. *The solution operators $\mathcal{S}_a, \mathcal{S}_s, \mathcal{S}_b : X \to Y$ in (10.1.1) and (10.1.2) are well-defined and belong to $\mathcal{B}_\psi(X;Y)$, where we assume for $\mathcal{S}_s, \mathcal{S}_b : X \to Y$ that $a > 0$ and $|\rho| < 1$.*

Proof. For the solution operator $\mathcal{S}_a : X \to Y$ solving the variational inequality in (10.2.1), we first observe that the continuous embeddings $H^1_\gamma(E) \subseteq L^2_\gamma(E) \cong$

$L^2_\gamma(E)^* \subseteq H^1_\gamma(E)^*$ hold true such that $(H^1_\gamma(E), L^2_\gamma(E), H^1_\gamma(E)^*)$ is a Gelfand triple, where $H^1_\gamma(E) := \left\{f : E \to \mathbb{R};\, f, \partial^\alpha f \in L^2_\gamma(E),\ \text{for all } |\alpha| \leqslant 1\right\}$. Then, it follows by taking the limit of the discretization in time as in [HRSW13, Theorem B.2.2] that the map $\mathcal{S}_a : L^2_\gamma(E) \to L^2_\gamma([0,T] \times E \times F)$ is bounded and Lipschitz continuous with respect to the norm topology of $L^2_\gamma(E)$, where we use that the interest rate is constant. Since the weak-$*$-topology is weaker than the norm topology, it follows that $\mathcal{S}_a \in C^0_b(X;Y) \subset \mathcal{B}_\psi(X;Y)$.

For the other operators $\mathcal{S}_s, \mathcal{S}_b : X \to Y$, we first show that $\mathcal{S}_s : X \to Y$ satisfies a Lipschitz estimate, from which it follows that $\mathcal{S}_s \in C^0_b(X;Y) \subset \mathcal{B}_\psi(X;Y)$, where we conclude by changing sign that also $\mathcal{S}_b \in C^0_b(X;Y) \subset \mathcal{B}_\psi(X;Y)$. We fix two functions $f, g \in X$ representing different terminal conditions and assume that $u, v : [0,T]\times]0,\infty[\times[a,b] \to \mathbb{R}$ are the corresponding viscosity solutions of (9.4.6), which exist by Theorem 9.9. Furthermore, we observe that the logarithmic stock price $\log X_0(T)$ starting at zero at time $t \in [0,T]$ under the EMM $\mathbb{Q}_\lambda \sim \mathbb{P}$, conditional on $(\sigma(s))_{t\leqslant s\leqslant T}$, is Gaussian with mean $M(t) := \int_t^T \left(r - \sigma(s)^2/2\right) ds + \frac{\rho}{\xi}\left(\sigma(T) - \sigma(t) - \int_t^T (\kappa\theta - \lambda - \kappa\sigma(s))ds\right)$ and variance $C(t) := \int_t^T \left(\sigma(s) - \rho^2 Q(\sigma(s))\right) ds$. The density of this conditional random variable under $\mathbb{Q}_\lambda \sim \mathbb{P}$ is of the form

$$p^{t,\sigma(t),\lambda}_{\log X_0|\sigma}(z; M(t), C(t)) = \frac{1}{\sqrt{2\pi C(t)}} \exp\left(-\frac{(z - M(t))^2}{2C(t)}\right),$$

for $z \in \mathbb{R}$, such that $p_{t,y,\lambda}(z) = \mathbb{E}^{\mathbb{Q}_\lambda}\left[p^{t,\sigma(t),\lambda}_{\log X_0|\sigma}(z; M(t), C(t))\middle|\sigma(t) = y\right]$ is by Fubini's theorem measurable and satisfies for every bounded, measurable function $f : \mathbb{R} \to \mathbb{R}$ that

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}_\lambda}\left[f(\log X(T))|X(t) = x, \sigma(t) = y\right] \\
&\quad = \mathbb{E}^{\mathbb{Q}_\lambda}\left[\int_{\mathbb{R}} f(z + \log x) p^{t,y,\lambda}_{\log X_0|\sigma}(z; M(t), C(t))dz \middle| X(t) = 1, \sigma(t) = y\right] \\
&\quad = \int_{\mathbb{R}} f(y + \log x) p_{t,y,\lambda}(z)dz, \quad \mathbb{P}\text{-a.s.}
\end{aligned}$$

Moreover, $a > 0$ and $|\rho| < 1$ imply $\mathbb{E}\left[C(t)^{-1/2-k}\right] < \infty$, for all $k \in \mathbb{N}$, such that dominated convergence implies that $z \mapsto e^{\varepsilon z^2} p_{t,y,\lambda}(z)$ with $\varepsilon < 1/(2bT)$ is uniformly bounded, see [AFP18, Theorem 3.1]. From this, we define $p(z) = \sup_{t\in[0,T], y\in[a,b], \lambda\in\Lambda} p_{t,y,\lambda}(z)$ such that $z \mapsto e^{\varepsilon z^2} p(z)$ is also uniformly bounded as supremum of uniformly bounded functions over the compact subset $[0,T] \times [a,b] \times \overline{\Lambda}$. Using the definition of $u_s : [0,T]\times]0,\infty[\times[a,b] \to \mathbb{R}$ in (9.4.5), we

conclude by the substitution $\eta(x,z) = (x\exp(z), z)$ that

$$\begin{aligned}
\|u-v\|^2_{L^2_\gamma([0,T]\times]0,\infty[\times[a,b])} &\leqslant \int_0^T \int_0^\infty \int_a^b \gamma(x)^2 |u(t,x,y) - v(t,x,y)|^2 dydxdt \\
&\leqslant \int_0^T \int_0^\infty \int_a^b \gamma(x)^2 e^{-2r(T-t)} \cdots \\
&\quad \cdots \operatorname*{ess\,sup}_{\lambda\in\Lambda} \mathbb{E}^{\mathbb{Q}_\lambda} \left[|f(X(T)) - g(X(T))|^2 \,\middle|\, (X(t),\sigma(t)) = (x,y) \right] dydxdt \\
&\leqslant (b-a)\frac{1-e^{-2rT}}{2r} \int_0^\infty \gamma(x)^2 \cdots \\
&\quad \cdots \sup_{t\in[0,T],\lambda\in\mathbb{R},y\in[a,b]} \int_{\mathbb{R}} \left| f\left(xe^z\right) - g\left(xe^z\right) \right|^2 p_{t,y,\lambda}(z) dzdx \\
&\leqslant (b-a)\frac{1-e^{-2rT}}{2r} \sup_{z\in\mathbb{R}} \left(e^{\varepsilon z^2} p(z) \right) \int_0^\infty \gamma(x)^2 \cdots \\
&\quad \cdots \int_{\mathbb{R}} e^{-\varepsilon z^2} \left| f\left(xe^z\right) - g\left(xe^z\right) \right|^2 dzdx \\
&\leqslant (b-a)\frac{1-e^{-2rT}}{2r} \sup_{z\in\mathbb{R}} \left(e^{\varepsilon z^2} p(z) \right) C \|f-g\|^2_{L^2_\gamma(]0,\infty[)},
\end{aligned}$$

where $C = \int_{\mathbb{R}} \exp\left(-\varepsilon z^2 - z\right) dz = \sqrt{\pi e^{1/(2\varepsilon)}/\varepsilon} < \infty$. By choosing the zero payoff function $g = 0$ such that $v = 0$, it follows that $\mathcal{S}_s : X \to Y$ has bounded operator norm, whereas by inserting a sequence $(f_n)_{n\in\mathbb{N}}$ in X converging to $g \in X$ in $L^2_\gamma(]0,\infty[)$, we conclude that $\mathcal{S}_s(f_n) \to \mathcal{S}_s(g)$ in $L^2_\gamma([0,T]\times]0,\infty[\times[a,b])$, as $n \to \infty$. Because the weak-$*$-topology is a weaker topology than the norm topology on $L^2_\gamma(]0,\infty[)$, it follows that $\mathcal{S}_s \in C^0_b(X;Y) \subset \mathcal{B}_\psi(X;Y)$, which holds by similar arguments also true for $\mathcal{S}_b : X \to Y$. □

In the following, we approximate the solution operators $\mathcal{S}_a, \mathcal{S}_s, \mathcal{S}_b : X \to Y$ by an abstract neural network. For this purpose, we construct an additive family $\mathcal{H}$ between the input space (X,ψ) and the hidden layer space $Z = \mathbb{R}$ with admissible weight function $\psi_Z(z) = \exp(|z|)$. Let $\mu_1, ..., \mu_m \in \mathbb{R}$ and $\tau_1, ..., \tau_m \in \mathbb{R}$ be constants, from which we define $\sigma_j = \exp(\tau_j)$ to account for the non-negativity of the variance. Using the probability density functions of the normal distribution, we define

$$p_j(w) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(w-\mu_j)^2}{2\sigma_j^2} \right),$$

for $w \in \mathbb{R}$ and $j = 1, ..., m$. Using them as kernels, we define the additive family

$\mathcal{H}$ as hidden layer maps $h : X \to \mathbb{R}$ of the form

$$h(g) = c_0 + \sum_{j=1}^{m} c_j \int_{\mathbb{R}} g(w)p_j(w)dw, \tag{10.1.3}$$

for $g \in X$ and some $c_0, ..., c_m \in \mathbb{R}$, where $g : E \to \mathbb{R}$ is extended to $\mathcal{R}$ by setting it to zero outside of E. Then, $\mathcal{H}$ is closed under addition and as $p_j : \mathbb{R} \to \mathbb{R}$ are probability density function of normal distributed random variable, we have

$$h(g) = c_0 + \sum_{j=1}^{m} c_j \mathbb{E}\left[g(Z_j)\right],$$

with $Z_j \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right)$. Hence, a hidden layer map $h : X \to \mathbb{R}$ can be seen as linear combinations of option prices with payoff function $g : E \to \mathbb{R}$ if the underlying process X_j is assumed to be a Brownian motion over the time interval $[0, 1]$ with drift μ_j and variance σ_j, i.e.

$$X_j(t) = \mu_j t + \sigma_j B_j(t),$$

for $t \in [0, 1]$, such that $Z_j = X_j(1)$, where $(B_j)_{j=1,...,m}$ are m independent standard Brownian motions. In this case, the expected value of the European call and put options with payoff functions $g_1(x) = \max(x - K, 0)$ and $g_2(x) = \max(K - x, 0)$ can be computed with the Bachelier option pricing formula such that

$$\begin{aligned}
\mathbb{E}\left[g_1(Z_j)\right] &= \int_K^\infty \frac{x-K}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} dx \\
&= (\mu_j - K)\Phi_{\mathcal{N}(0,1)}\left(\frac{\mu_j - K}{\sigma_j}\right) + \sigma_j \varphi_{\mathcal{N}(0,1)}\left(\frac{\mu_j - K}{\sigma_j}\right), \\
\mathbb{E}\left[g_2(Z_j)\right] &= \int_{-\infty}^K \frac{K-x}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} dx \\
&= (K - \mu_j)\Phi_{\mathcal{N}(0,1)}\left(\frac{K - \mu_j}{\sigma_j}\right) + \sigma_j \varphi_{\mathcal{N}(0,1)}\left(\frac{K - \mu_j}{\sigma_j}\right),
\end{aligned}$$

for $j = 1, ..., m$, where $\varphi_{\mathcal{N}(0,1)} : \mathbb{R} \to \mathbb{R}$ and $\Phi_{\mathcal{N}(0,1)} : \mathbb{R} \to [0, 1]$ denote the probability density function and the cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$, respectively. Similarly, the expected values of the digital European call and put cash-or-nothing options with payoff

functions $g_3(x) = \mathbb{1}_{\{x \geqslant K\}}$ and $g_4(x) = \mathbb{1}_{\{x < K\}}$ are given by

$$\mathbb{E}\left[g_3(Z_j)\right] = \int_K^\infty \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} dx = \Phi_{\mathcal{N}(0,1)}\left(\frac{\mu_j - K}{\sigma_j}\right),$$
$$\mathbb{E}\left[g_4(Z_j)\right] = \int_{-\infty}^K \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} dx = \Phi_{\mathcal{N}(0,1)}\left(\frac{K - \mu_j}{\sigma_j}\right),$$

for $j = 1, ..., m$. This means that the hidden layer map $h : X \to \mathbb{R}$ can be computed in closed form for every of these four options such that an abstract neural network can be efficiently implemented on a computer. However, the expected values are also known for other popular options in a Bachelier model, see for example [Hau98], which can be generalized to multi-dimensional options written on several underlying assets.

Using the additive family in (10.1.3), we construct an abstract neural network similarly as in in Example 3.19 by choosing the activation family $\mathcal{R} = \{\rho\}$ and the linear readout $\mathcal{L} = \left\{z \mapsto \phi z;\, \phi \in \mathcal{NN}_{X_0,\mathbb{R}}^{\mathcal{G},\{\rho_1\},\mathcal{K}}\right\}$ with $X_0 = [0,T] \times E \times F$, where the activation function $\rho_1 \in L^2_\gamma(\mathbb{R})$ is chosen such that $\mathcal{NN}_{X_0,\mathbb{R}}^{\mathcal{G},\{\rho_1\},\mathcal{K}}$ is dense in Y, and where $\mathcal{G} = \left\{X_0 \ni (t,x,y) \mapsto (t,x,y)a + b \in \mathbb{R};\, a \in \mathbb{R}^3,\, b \in \mathbb{R}\right\}$ and $\mathcal{K} = \{\mathbb{R} \ni z \mapsto wz \in \mathbb{R};\, w \in \mathbb{R}\}$, which implies $\beta(z) \in L^2_\gamma([0,T] \times E \times F)$, for all $z \in \mathbb{R}$. Hence, an abstract neural network is of the form

$$\varphi(g) = \sum_{k=1}^N \phi_k \rho(h_k(g)) = \left((t,x,y) \mapsto \sum_{k=1}^N \phi_k(t,x,y)\rho(h_k(g))\right), \qquad (10.1.4)$$

for $g \in X$ and some $N \in \mathbb{N}_{\geqslant 1}$, where $\phi_k \in \mathcal{NN}_{X_0,\mathbb{R}}^{\mathcal{G},\{\rho_1\},\mathcal{K}}$ are conventional neural networks, $h_1, ..., h_N \in \mathcal{H}$ are the hidden layer maps, and $\rho : \mathbb{R} \to \mathbb{R}$ is an activation function such that $\mathcal{NN}_{\mathbb{R},\mathbb{R}}^{\mathcal{G}_0,\{\rho\},\mathcal{K}_0}$ is $\tau^0_{\psi_\mathbb{R}}$-dense in $\mathcal{B}_{\psi_\mathbb{R}}(\mathbb{R})$, where $\mathcal{G}_0 = \{\mathbb{R} \ni z \mapsto az + b \in \mathbb{R};\, a, b \in \mathbb{R}\}$ and $\mathcal{K}_0 = \{\mathbb{R} \ni z \mapsto wz \in \mathbb{R};\, w \in \mathbb{R}\}$.

Theorem 10.2. *For every $\varepsilon > 0$ there exists some $\varphi : X \to Y$ of the form (10.1.4) such that*

$$\|\mathcal{S} - \varphi\|_{\mathcal{B}_\psi(X;Y)} = \sup_{g \in X} \frac{\|\mathcal{S}(g) - \varphi(g)\|_Y}{\psi(g)} < \varepsilon,$$

where $\mathcal{S} : X \to Y$ is either the solution operator $\mathcal{S}_a : X \to Y$ of the American option price, the solution operator $\mathcal{S}_s : X \to Y$ of the seller's price, or the solution operator $\mathcal{S}_b : X \to Y$ of the buyer's price.

Proof. The solution operator $\mathcal{S} : X \to Y$ belongs by Lemma 10.1 to the space

$\mathcal{B}_\psi(X;Y)$ such that we are able to apply the universal approximation result in Theorem 3.27. First, we show that $\mathcal{H}$ defined in (10.1.3) is an additive family, where it follows by construction that $\mathcal{H}$ is closed under addition and contains the constant functions. To show that $\mathcal{H}$ is point separating, let $g_1, g_2 \in X$ be two distinct payoff functions, which differ on a measurable set A with $\int_A \gamma(x)dx > 0$, such that it follows for suitably chosen $\mu \in \mathbb{R}$ and $\sigma > 0$ that

$$h(g_1) = \int_0^\infty g_1(w)p(w)dw \neq \int_0^\infty g_2(w)p(w)dw = h(g_2),$$

where $p : \mathbb{R} \to \mathbb{R}$ denotes the probability density function of $\mathcal{N}\left(\mu, \sigma^2\right)$, which shows that $\mathcal{H}$ is point separating. Secondly, we check the assumptions of Theorem 3.27, where we consider the architectures $(\mathcal{H}, \mathcal{R}, \mathcal{L})$ between X and Y as well as $(\mathcal{G}_0, \mathcal{R}, \mathcal{K}_0)$ between $\mathbb{R}$ and $\mathbb{R}$ described above. Then, it follows for every $g \mapsto h(g) = c_0 + \sum_{j=1}^m c_j \int_0^\infty g(w)p(w)dw$ and $z \mapsto g_0(z) = az + b \in \mathcal{G}$ that

$$g \mapsto (g_0 \circ h)(g) = (ac_0 + b) + \sum_{j=1}^m ac_j \int_0^\infty g(w)p(w)dw \in \mathcal{H}.$$

Moreover, we conclude for every $g \mapsto h(g) := \int_0^\infty g(w)p(w)dw$ by the Cauchy-Schwarz inequality that

$$\begin{aligned}\sup_{g\in X} \frac{\psi_Z(h(g))}{\psi(g)} &= \sup_{g\in X} \frac{\exp\left(\left|\int_0^\infty g(w)p(w)dw\right|\right)}{\exp\left(\|g\|^2_{L^2_\gamma(]0,\infty[)}\right)} \\ &\leqslant \sup_{g\in X} \frac{\exp\left(C_0\|g\|_{L^2_\gamma(]0,\infty[)}\right)}{\exp\left(\|g\|^2_{L^2_\gamma(]0,\infty[)}\right)} < \infty,\end{aligned}$$

with $C_0 = \left(\int_{\mathbb{R}} \gamma(w)^2 p(w)^2 dx\right)^{1/2} < \infty$, which shows by taking linear combinations that $\sup_{g\in X} \frac{\psi_Z(h(g))}{\psi(g)} < \infty$, for all $h \in \mathcal{H}$. Moreover, we have $\Gamma = \{\gamma \in L(\mathbb{R};Y);\, \forall \alpha \in \mathcal{K}_0 : \gamma \circ \alpha \in \mathcal{L}\} = \mathcal{L}$ such that $\bigcup_{\gamma\in\Gamma,\, z\in\mathbb{R}} \{\gamma(z)\} = \mathcal{NN}^{\mathcal{G},\{\rho_1\},\mathcal{K}}_{X_0,\mathbb{R}}$ is by assumption dense in Y. Hence, $\mathcal{NN}^{\mathcal{H},\mathcal{R},\mathcal{L}}_{X,Y}$ is by Theorem 3.27 τ^0_ψ-dense in $\mathcal{B}_\psi(X;Y)$. □

Theorem 10.2 provides the theoretical basis to use abstract neural networks for learning the solution operator of the non-linear HJB equation, where there exists a large enough abstract neural network of the form (10.1.4) such that the solution operator $\mathcal{S}_a, \mathcal{S}_s, \mathcal{S}_b : X \to Y$ is approximated well enough.

10.2 Learning American Options

Given the prices of several American call and put options on a single asset, we want to learn the solution operator of the corresponding HJB equation. We assume that X follows under the real-world measure $\mathbb{P}$ a geometric Brownian motion, i.e.

$$dX(t) = \sigma X(t) dB(t),$$

for $t \in [0,T]$, with initial value $X(0) > 0$, where $\sigma = 0.3$ is the volatility parameter. Hence, the value function describing an European option with payoff function $g :]0,\infty[\to \mathbb{R}$ is given by

$$u(t,x) = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^{\mathbb{Q}}\left[e^{-r(T-\tau)} g(X(\tau)) \middle| X(t) = x \right], \quad \mathbb{Q}\text{-a.s.},$$

for $(t,x) \in]0,\infty[$, where $r = 0.01$. In this case, $u : [0,T] \times]0,\infty[\to \mathbb{R}$ is a solution to the HJB variational inequality

$$\max\left(\partial_t u(t,x) + \frac{1}{2} x^2 \partial_{xx} u(t,x) - r u(t,x), g(x) - u(t,x) \right) = 0, \qquad (10.2.1)$$

for all $(t,x) \in [0,T[\times]0,\infty[$, with terminal condition $u(T,x) = g(x)$, for all $x \in]0,\infty[$.

In order to learn the solution operator of the HJB variational inequality in (10.2.1), the prices $u(t,x) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^{\mathbb{Q}}\left[e^{-r(T-\tau)} g_{1,2}(X(\tau)) \middle| X(t) = x \right]$ of the American call option $g_1(x) = \max(x - K, 0)$ and the American put option $g_2(x) = \max(K - x, 0)$ are assumed to be known, where we add the initial price $g_3(x) = x$ satisfying $\mathcal{S}_a(g_3)(t,x) = e^{-r(T-t)} x$ and the strike level $g_4(x) = K$ satisfying $\mathcal{S}_a(g_4)(t,x) = e^{-r(T-t)} K$ as constraints. More precisely, we assume that the prices $u(t,x)$ are known for the payoff functions $g_1, ..., g_4 :]0,\infty[\to \mathbb{R}$, times $t = 1/12, 2/12, ..., 12/12$ (representing months), initial values $x = 0.5, 1, 1.5, ..., 100$ and strike prices $K = 0.5, 1, 1.5, ..., 100$, which yields $4 \cdot 12 \cdot 200 \cdot 200 = 1.92 \cdot 10^6$ training data points. From this, we predict the American digital call and put option prices given by $u(t,x) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^{\mathbb{Q}}\left[e^{-r(T-\tau)} g_{5,6}(X(\tau)) \middle| X(t) = x \right]$, with the payoff functions $g_5(x) = \mathbb{1}_{\{x \geqslant K\}}$ and $g_6(x) = \mathbb{1}_{\{x < K\}}$. The abstract neural networks are built in `Python` using `Tensorflow`, where we choose the size $(N, m, N_1) = (40, 20, 20)$ such that $5'640$ parameters need to be estimated. For training, the Adam algorithm is applied over $100'000$ epochs with a learning rate of 10^{-4}. On the other hand, the true data is calculated in `RStudio` with the library `RQuantLib` of the open-source software `QuantLib`, where the functions `AmericanOption` and `BinaryOption` are used to compute the prices of the digital options. The results are plotted in Figure 10.1.

10.3 Learning Super-Replication Price with Stochastic Volatility

In this section, we learn the solution operator of the HJB equation corresponding to the super-replication and sub-replication problem of an European option written on a risky asset following a Jacobi stochastic volatility model. More precisely, we assume that X and σ follow under the real world measure $\mathbb{P}$ the SDE

$$\begin{aligned} dX(t) &= rX(t)dt + \rho X(t)\sqrt{Q(\sigma(t))}dB_1(t) + X(t)\sqrt{\sigma(t) - \rho^2 Q(\sigma(t))}dB_2(t), \\ d\sigma(t) &= \kappa(\theta - \sigma(t))dt + \xi\sqrt{Q(\sigma(t))}dB_1(t), \end{aligned}$$

for $t \in [0,T]$, with initial values $X(0) \in [1,100]$ and $\sigma(0) \in [0.01, 0.4]$, where $Q(v) = \frac{(b-v)(v-a)}{(\sqrt{b}-\sqrt{a})^2}$ with $a = 0.01$, $b = 0.4$, and where $r = 0.01$, $\kappa = 1$, $\theta = 0.2$, $\xi = 0.5$ and $\rho = -0.5$. Using Girsanov's theorem to find equivalent martingale measures as in Section 9.4, the super-replication price (seller's price) and sub-replication price (buyer's price) of an European option with payoff function $g :]0,\infty[\to \mathbb{R}$ are given by

$$\begin{aligned} u_s(t,x,y) &= \operatorname*{ess\,sup}_{\lambda\in\Lambda} \mathbb{E}^{\mathbb{Q}_\lambda}\left[e^{-r(T-t)}g(X(T))\Big|X(t) = x, \sigma(t) = y\right], \\ u_b(t,x,y) &= \operatorname*{ess\,inf}_{\lambda\in\Lambda} \mathbb{E}^{\mathbb{Q}_\lambda}\left[e^{-r(T-t)}g(X(t))\Big|X(t) = x, \sigma(t) = y\right], \end{aligned} \tag{10.3.1}$$

$\mathbb{P}$-a.s., for $(t,x,y) \in [0,T]\times]0,\infty[\times[a,b]$, where $\Lambda = [\kappa(\theta-b), \kappa(\theta-a)[$. In this case, $u_s, u_b : [0,T]\times]0,\infty[\to\mathbb{R}$ are solutions of

$$\begin{aligned} -\partial_t u_s(t,x,y) - \sup_{\lambda\in\Lambda}\left(\left(\mathcal{A}^\lambda u_s\right)(t,x,y) - r u_s(t,x,y)\right) = 0, \\ -\partial_t u_b(t,x,y) - \inf_{\lambda\in\Lambda}\left(\left(\mathcal{A}^\lambda u_b\right)(t,x,y) - r u_b(t,x,y)\right) = 0, \end{aligned} \tag{10.3.2}$$

for all $(t,x) \in [0,T]\times]0,\infty[\times[a,b]$, with terminal conditions $u_{s,b}(T,x) = g(x)$, for all $x \in]0,\infty[$, where

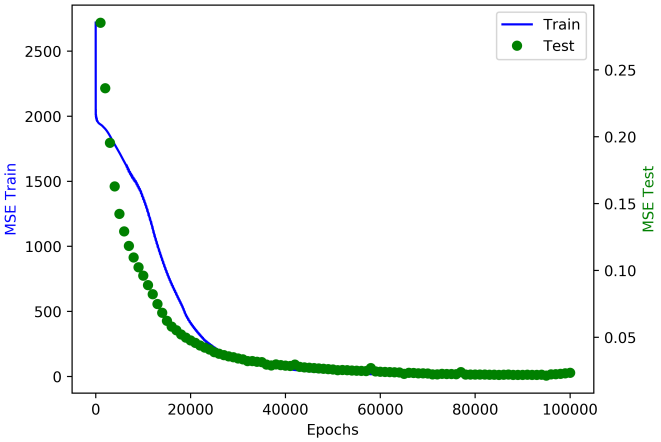
$$\begin{aligned} \left(\mathcal{A}^\lambda f\right)(x,y) = \frac{1}{2}x^2 y\partial_{xx}f(x,y) + \rho\xi x Q(y)\partial_{xy}f(x,y) + \frac{1}{2}\xi^2 Q(y)\partial_{yy}f(x,y) \\ + rx\partial_x f(x,y) + (\kappa\theta - \lambda - \kappa y)\partial_y f(x,y), \end{aligned} \tag{10.3.3}$$

for $f \in C^2(]0,\infty[\times[a,b])$ and $x \in]0,\infty[\times[a,b]$.

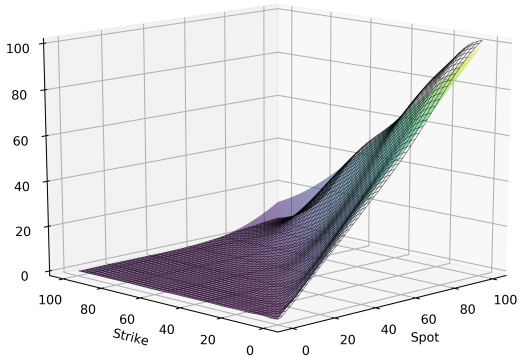
In order to learn the solution operators, we assume that the super- and sub-replication prices $u_s(t,x,y)$ resp. $u_b(t,x,y)$ of the European call option $g_1(x) = \max(x-K,0)$ and the European put option $g_2(x) = \max(K-x,0)$

are known, where we add as constraints the initial price $g_3(x) = x$ satisfying $\mathcal{S}_s(g_3)(t,x,y) = \sup_{\lambda\in\Lambda} \mathbb{E}^{\mathbb{Q}_\lambda}\left[e^{-r(T-t)}X(t)\big|(X(t),\sigma(t)) = (x,y)\right]$ (resp. with the infimum over $\lambda \in \Lambda$ for $\mathcal{S}_b : X \to Y$ instead of the supremum) and the strike level $g_4(x) = K$ satisfying $\mathcal{S}_{s,b}(g_4)(t,x,y) = e^{-r(T-t)}K$. More precisely, we assume that the prices $u_{s,b}(t,x,y)$ are known for the payoff functions $g_1, ..., g_4 :]0,\infty[\to \mathbb{R}$, times $t = 1/12, 2/12, ..., 1$ (representing months), initial values $x = 1, 2, ..., 100$ and $y = 0.01, 0.015, ..., 0.5$, and strike prices $K = 1, 2, ..., 100$, which yields $4 \cdot 12 \cdot 100 \cdot 79 \cdot 100 = 3.792 \cdot 10^7$ training data points. The abstract neural networks are built in `Python` using `Tensorflow`, where we choose the size $(N, m, N_1) = (40, 20, 20)$ such that $6'440$ parameters need to be estimated. For training, the Adam algorithm is applied over $100'000$ epochs with a learning rate of 10^{-4}. The results are plotted in Figure 10.2 and Figure 10.3.

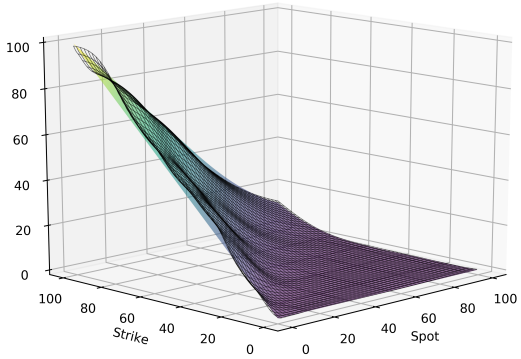
Hereby, the true super-replication and sub-replication prices of the European call, put, digital call, and digital put options are computed as in [AFP18], where we first observe that the supremum or infimum in (10.3.1) resp. (10.3.2) is attained either for $\lambda = \kappa(\theta - b)$ or $\lambda = \kappa(\theta - a)$, because λ appears linearly in (10.3.3). Then, the expected value under $\mathbb{Q}_\lambda$ is approximated using a truncated series of the Gram-Charlier expansion of Type A for the distribution of the joint distribution $(\log X(T), \sigma(T))$, whose moments can be computed explicitly with the help of the polynomial property in [FL16, Theorem 3.1] and the payoff is transformed with Fourier techniques, see [AFP18, Theorem 3.7+3.8] and [AFP18, Appendix A].



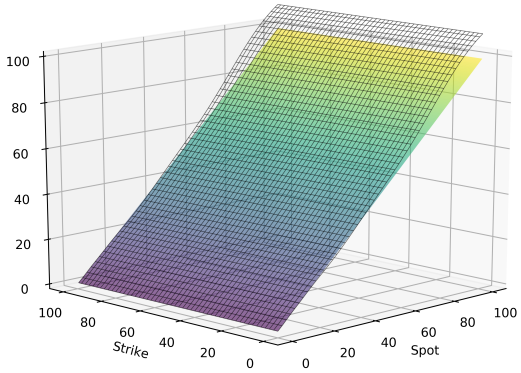
(a) Learning performance



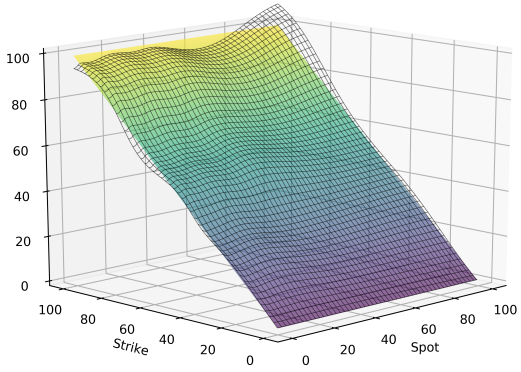
(b) American call option $g_1(x) = (x - K)_+$



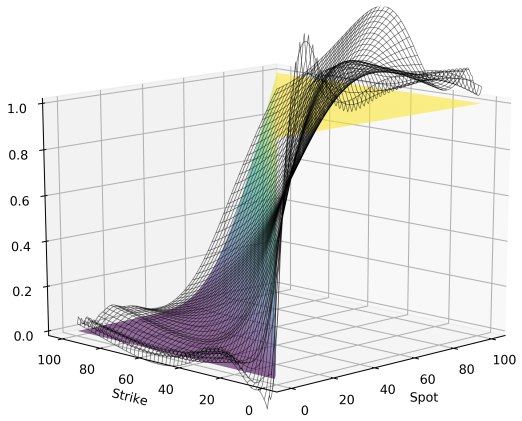
(c) American put option $g_2(x) = (K - x)_+$



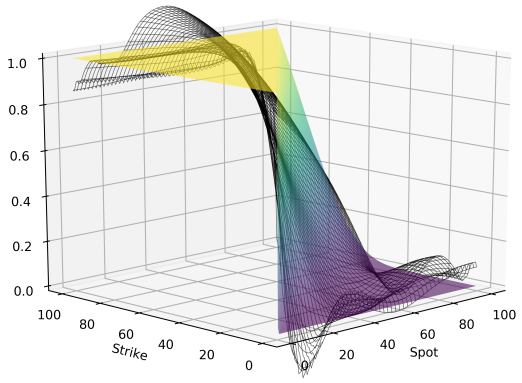
(d) Initial price constraint $g_3(x) = x$



(e) Strike level constraint $g_4(x) = K$

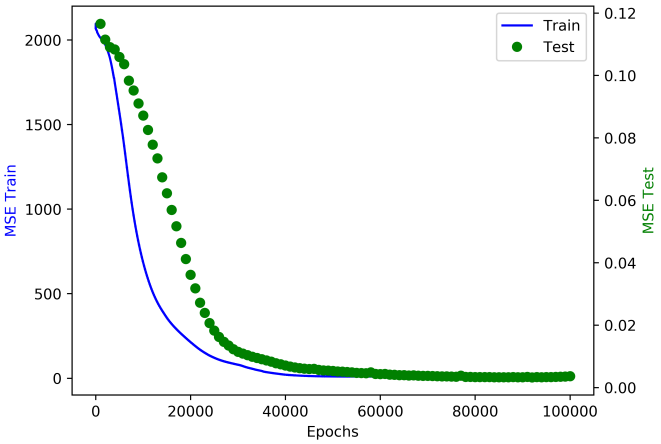


(f) American digital call option $g_5(x) = \mathbb{1}_{\{x \geqslant K\}}$

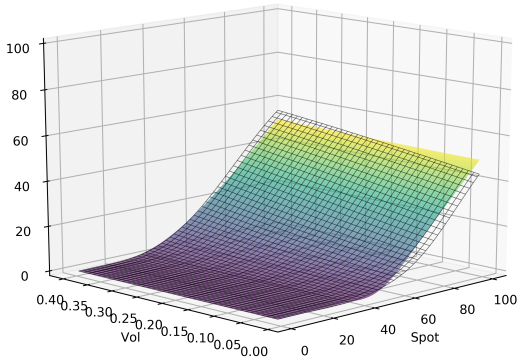


(g) American digital put option $g_6(x) = \mathbb{1}_{\{x < K\}}$

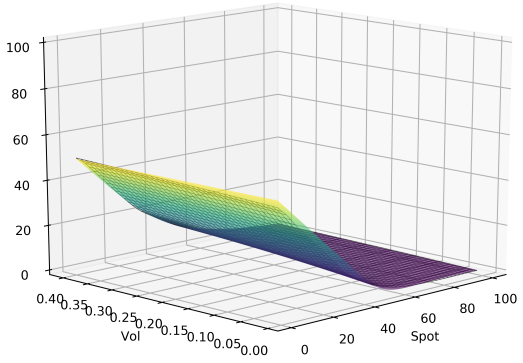
Figure 10.1: Learning the solution operator $\mathcal{S}_a : X \to Y$ of the American option price. In (a), the loss during the training procedure is displayed. In (b) and (c), the true data (surface) and the learned value functions (wireframes) of the American call and put options at time $t = 1$ are shown, while the constraints (initial price and strike level) are plotted in (d) and (e). In (f) and (g), the predicted value functions (wireframes) of the American digital call and put options at time $t = 1$ are compared to the true data (surface).



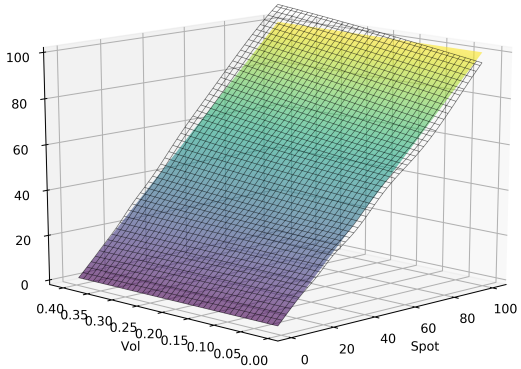
(a) Learning performance



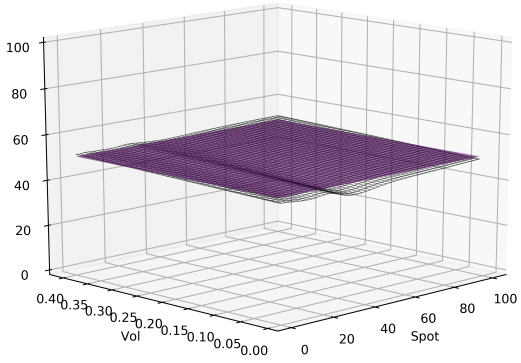
(b) European call option $g_1(x) = (x - K)_+$



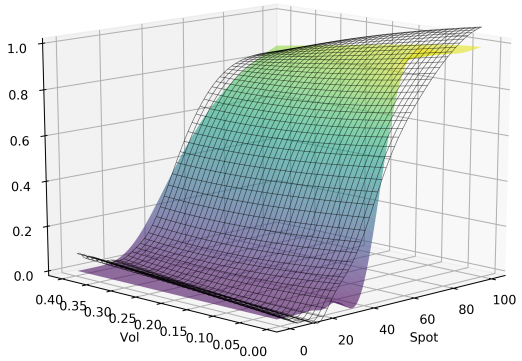
(c) European put option $g_2(x) = (K - x)_+$



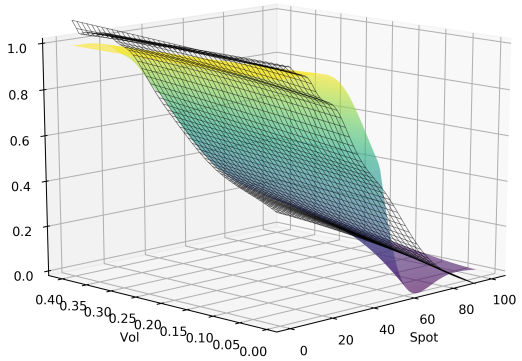
(d) Initial price constraint $g_3(x) = x$



(e) Strike level constraint $g_4(x) = K$

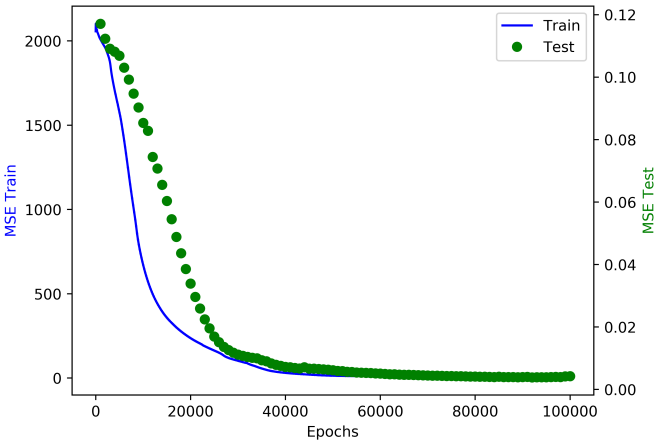


(f) European digital call option $g_5(x) = \mathbb{1}_{\{x \geqslant K\}}$

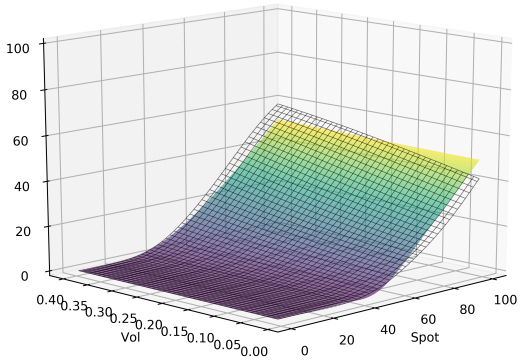


(g) European digital put option $g_6(x) = \mathbb{1}_{\{x < K\}}$

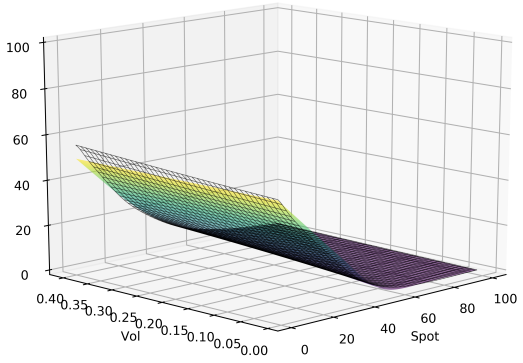
Figure 10.2: Learning the solution operator $\mathcal{S}_s : X \to Y$ of the super-replication (seller's) price. In (a), the loss during the training procedure is displayed. In (b) and (c), the true data (surface) and the learned value functions (wireframes) of the European call and put options at time $t = 1$ are shown, while the constraints (initial price and strike level) are plotted in (d) and (e). In (f) and (g), the predicted value functions (wireframes) of the European digital call and put options at time $t = 1$ with $K = 50$ are compared to the true data (surface).



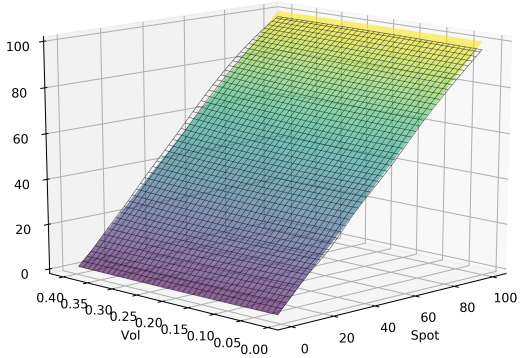
(a) Learning performance



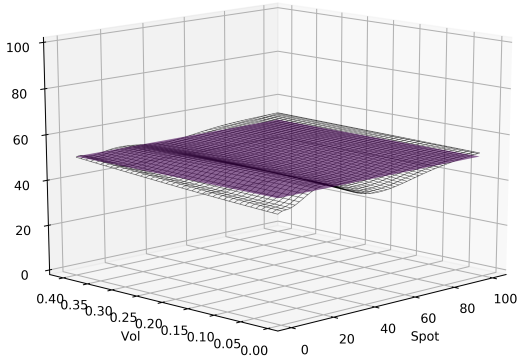
(b) European call option $g_1(x) = (x - K)_+$



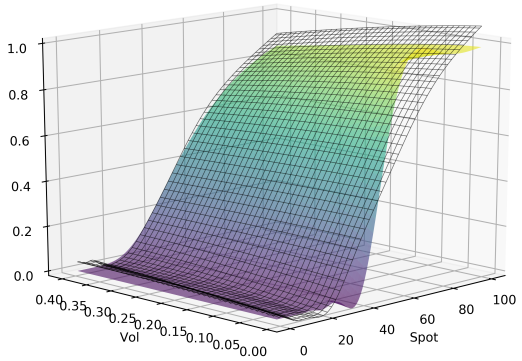
(c) European put option $g_2(x) = (K - x)_+$



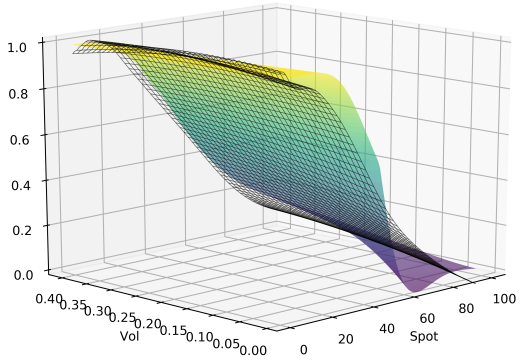
(d) Initial price constraint $g_3(x) = x$



(e) Strike level constraint $g_4(x) = K$



(f) European digital call option $g_5(x) = \mathbb{1}_{\{x \geqslant K\}}$



(g) European digital put option $g_6(x) = \mathbb{1}_{\{x < K\}}$

Figure 10.3: Learning the solution operator $\mathcal{S}_b : X \to Y$ of the sub-replication (buyer's) price. In (a), the loss during the training procedure is displayed. In (b) and (c), the true data (surface) and the learned value functions (wireframes) of the European call and put options at time $t = 1$ are shown, while the constraints (initial price and strike level) are plotted in (d) and (e). In (f) and (g), the predicted value functions (wireframes) of the European digital call and put options at time $t = 1$ with $K = 50$ are compared to the true data (surface).

Bibliography

Bibliography

[AFP18] Damien Ackerer, Damir Filipović, and Sergio Pulido. "The Jacobi stochastic volatility model". In: *Finance and Stochastics* 22 (3 2018), pp. 667–700.

[Ada95] E. Adam. "Approximation properties and spaces of smooth functions". PhD thesis. Vienna: Universität Wien, 1995.

[Ada76] David R. Adams. "On the existence of capacitary strong type estimates in $\mathbb{R}^n$". In: *Arkiv för Matematik* 14.1-2 (1976), pp. 125–140.

[AF08] Robert A. Adams and John F. Fournier. *Sobolev Spaces*. 2nd ed., [Repr.] Pure and applied mathematics series 140, repr. 2008. Amsterdam: Academic Press, 2008.

[AB06] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. 3rd ed. Berlin, Heidelberg: Springer, 2006.

[AGLM93] Luis Alvarez et al. "Axioms and fundamental equations of image processing". In: *Archive for Rational Mechanics and Analysis* 123 (3 1993), pp. 199–257.

[AC17] Anna Ananova and Rama Cont. "Pathwise integration with respect to paths of finite quadratic variation". In: *Journal de Mathématiques Pures et Appliquées* 107.6 (2017), pp. 737–757.

[Apo67] Tom M. Apostol. *Calculus*. Vol. 1. New York: John Wiley & Sons, 1967.

[AP80] Richard M. Aron and João B. Prolla. "Polynomial approximation of differentiable functions on Banach spaces." In: *Journal für die reine und angewandte Mathematik* 313 (1980), pp. 195–216.

[Arr53] Kenneth J. Arrow. "Le rôle des valeurs boursières pour la répartition la meilleure des risques". In: *Économetrie*. Colloques Internationaux du Centre National de la Recherche Scientifique XI (1953), pp. 41–47.

[Ath19] Susan Athey. "The Impact of Machine Learning on Economics". In: *The Economics of Artificial Intelligence: An Agenda*. Ed. by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press, 2019, pp. 507–552.

[ASB20] Francesco Audrino, Fabio Sigrist, and Daniele Ballinari. "The impact of sentiment and attention measures on stock market volatility". In: *International Journal of Forecasting* 36.2 (2020), pp. 334–357.

[AD19] Saqib Aziz and Michael Dowling. "Machine Learning and AI for Risk Management". In: *Disrupting Finance: FinTech and Strategy in the 21st Century*. Ed. by Theo Lynn et al. Cham: Springer International Publishing, 2019, pp. 33–50.

[Bar94] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*. Mathématiques et applications 17. Paris: Springer, 1994.

[Bar93] Alexander R. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". In: *IEEE Transactions on Information Theory* 39.3 (1993), pp. 930–945.

[BS18] Christian Bayer and Benjamin Stemper. *Deep Calibration of Rough Stochastic Volatility Models*. 2018. arXiv: `1810.03399 [q-fin.PR]`.

[BCJ19] Sebastian Becker, Patrick Cheridito, and Arnulf Jentzen. "Deep Optimal Stopping". In: *Journal of Machine Learning Research* 20.74 (2019), pp. 1–25.

[BL82] Alain Bensoussan and Jacgues-Louis Lions. *Applications of variational inequalities in stochastic control*. Studies in mathematics and its applications; v. 12. Amsterdam: North-Holland Pub. Co., 1982.

[BDG21] Fred Espen Benth, Nils Detering, and Luca Galimberti. *Neural Networks in Fréchet spaces*. 2021. arXiv: `2109.13512 [math.FA]`.

[Ber24] Sergei Bernstein. "La problème de l'approximation des fonctions continues sur tout l'axe réel et l'une de ses applications". In: *Bulletin de la Société Mathématique de France* 52 (1924), pp. 399–410.

[Bis61] Errett Bishop. "A generalization of the Stone-Weierstrass theorem". In: *Pacific Journal of Mathematics* 11 (1961), pp. 777–783.

[BGLY16] Horatio Boedihardjo et al. "The signature of a rough path: Uniqueness". In: *Advances in Mathematics* 293 (2016), pp. 720–737.

[BGKP19] Helmut Bölcskei et al. "Optimal Approximation with Sparsely Connected Deep Neural Networks". In: *SIAM Journal on Mathematics of Data Science* 1 (2019), pp. 8–45.

[Bom67] Jan Boman. "Differentiability of a function and of its compositions with functions of one variable." In: *Mathematica Scandinavica* 20 (1967), pp. 249–268.

[Bou74] Nicolas Bourbaki. "Eléments de mathématique". In: Reprint of the 1974 original. Actualités Scientifiques et Industrielles. Berlin: Springer, 1974. Chap. 5-10.

[Bra07] Dietrich Braess. *Finite elements*. 3rd ed. Cambridge: Cambridge University Press, 2007.

[Bré11] Haïm Brézis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. New York: Springer, 2011.

[BD81] Bruno Brosowski and Frank Deutsch. "An Elementary Proof of the Stone-Weierstrass Theorem". In: *Proceedings of the American Mathematical Society* 81.1 (1981), pp. 89–92.

[Buc58] R. Creighton Buck. "Bounded continuous functions on a locally compact space". In: *Michigan Mathematical Journal* 5.2 (1958), pp. 95–104.

[BGTW19] Hans Buehler et al. "Deep Hedging". In: *Quantitative Finance* 19.8 (2019), pp. 1271–1291.

[Can98] Emmanuel J. Candès. "Ridgelets: Theory and Applications". PhD thesis. Stanford University, 1998.

[CJK16] Álvaro Cartea, Sebastian Jaimungal, and Damir Kinzebulatov. "Algorithmic Trading with Learning". In: *International Journal of Theoretical and Applied Finance* 19.04 (2016), pp. 1–30.

[CO18] Ilya Chevyrev and Harald Oberhauser. *Signature moments to characterize laws of stochastic processes*. 2018. arXiv: `1810.10971 [math.ST]`.

[CA84] Michael D. Cohen and Robert Axelrod. "Coping with Complexity: The Adaptive Value of Changing Utility". In: *The American Economic Review* 74.1 (1984), pp. 30–42.

[CS96] Gregory M. Constantine and Thomas H. Savits. "A Multivariate Faà Di Bruno Formula With Applications". In: *Transactions of the American Mathematical Society* 348.2 (1996), pp. 503–520.

[CF10] Rama Cont and David-Antoine Fournié. "Change of variable formulas for non-anticipative functionals on path space". In: *Journal of Functional Analysis* 259.4 (2010), pp. 1043–1072.

[CF13] Rama Cont and David-Antoine Fournié. "Functional Itô calculus and stochastic integral representation of martingales". In: *Annals of Probability* 41.1 (2013), pp. 109–133.

[CKT20] Christa Cuchiero, Wahid Khosrawi, and Josef Teichmann. "A generative adversarial network approach to calibration of local stochastic volatility models". In: *Risks* 8 (101 2020).

[CSW19] Christa Cuchiero, Walter Schachermayer, and Ting-Kam Leonard Wong. "Cover's universal portfolio, stochastic portfolio theory, and the numéraire portfolio". In: *Mathematical Finance* 29.3 (2019), pp. 773–803.

[CT20] Christa Cuchiero and Josef Teichmann. "Generalized Feller processes and Markovian lifts of stochastic Volterra processes: the affine case". In: *Journal of Evolution Equations* (2020), pp. 1–48.

[CT19] Christa Cuchiero and Josef Teichmann. "Markovian lifts of positive semidefinite affine Volterra-type processes". In: *Decisions in Economics and Finance* 42 (2019), pp. 407–448.

[CGOT20] Christa Cuchiero et al. *Discrete-time signatures and randomness in reservoir computing*. 2020. arXiv: 2010.14615 [cs.NE].

[Cyb89] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals and Systems* 2.4 (1989), pp. 303–314.

[DS94] Freddy Delbaen and Walter Schachermayer. "A general version of the fundamental theorem of asset pricing". In: *Mathematische Annalen* 300 (1 1994), pp. 463–520.

[DT10] Philipp Dörsek and Josef Teichmann. *A Semigroup Point Of View On Splitting Schemes For Stochastic (Partial) Differential Equations*. 2010. arXiv: 1011.2651.

[Dup09] Bruno Dupire. *Functional Itô Calculus*. Tech. rep. Bloomberg Portfolio Research Paper No. 2009-04-FRONTIERS. Bloomberg, 2009.

[Dup94] Bruno Dupire. "Pricing with a Smile". In: *Risk* 7 (1994), pp. 18–20.

[EHJ17] Weinen E, Jiequn Han, and Arnulf Jentzen. "Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations". In: *Communications in Mathematics and Statistics* 5 (4 2017), pp. 349–380.

[EKOO19] Sophie Emerson et al. "Trends and applications of machine learning in quantitative finance". In: *8th international conference on economics and finance research (ICEFR 2019)*. 2019.

[Eva15] Lawrence C. Evans. *Partial differential equations*. 2nd ed., repr. with corr. Graduate studies in mathematics vol. 19. Providence, R.I: American Mathematical Society, 2015.

[Faà57] Cavaliere Francesco Faà di Bruno. "Note sur une nouvelle formule de calcul différentiel". In: *The Quarterly Journal of Pure and Applied Mathematics* 1 (1857), pp. 359–360.

[Faà55] Cavaliere Francesco Faà di Bruno. "Sullo sviluppo delle Funzioni". In: *Annali di Scienze Matematiche e Fisiche* 6 (1855), pp. 479–480.

[Fer02] Robert E. Fernholz. *Stochastic Portfolio Theory*. 1st edition. Berlin Heidelberg: Springer Science + Business Media, 2002.

[FKK05] Robert Fernholz, Ioannis Karatzas, and Constantinos Kardaras. "Diversity and relative arbitrage in equity markets". In: *Finance and Stochastics* 9 (1 2005), pp. 1–27.

[FL16] Damir Filipović and Martin Larsson. "Polynomial diffusions and applications in finance". In: *Finance and Stochastics* 20.4 (2016), pp. 931–972.

[Föl81] Hans Föllmer. "Calcul d'Ito sans probabilités". In: *Séminaire de probabilités de Strasbourg* 15 (1981), pp. 143–150.

[FS03] Craig Friedman and Sven Sandow. "Learning Probabilistic Models: An Expected Utility Maximization Approach". In: *Journal of Machine Learning Research* 4 (2003), pp. 257–291.

[FH14] Peter K. Friz and Martin Hairer. *A Course on Rough Paths*. 1st ed. Universitext. Springer International Publishing, 2014.

[FV10] Peter K. Friz and Nicolas B. Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2010.

[FK88] Alfred Frölicher and Andreas Kriegl. *Linear spaces and differentiation theory*. Pure and applied mathematics. Chichester: J. Wiley, 1988.

[GTX19] Xiaojie Gao, Shikui Tu, and Lei Xu. *A* Tree Search for Portfolio Management*. 2019. arXiv: `1901.01855 [cs.AI]`.

[GK03] Gilles Godefroy and Nigel J. Kalton. "Lipschitz-free Banach spaces". In: *Studia Math* 159.1 (2003), pp. 121–141.

[GGO20] Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. *Approximation Bounds for Random Neural Networks and Reservoir Systems*. 2020. arXiv: `2002.05933 [math.PR]`.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[GO18] Lyudmila Grigoryeva and Juan-Pablo Ortega. "Echo state networks are universal". In: *Neural Networks* 108 (2018), pp. 495–508.

[GHJW18] Philipp Grohs et al. *A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations*. 2018. arXiv: `1809.02362 [math.NA]`.

[GH14] Julien Guyon and Pierre Henry-Labordère. *Nonlinear option pricing*. Chapman & Hall/CRC financial mathematics series. Boca Raton: CRC Press, 2014.

[Har08] Stevan Harnad. "The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence". In: *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Ed. by Robert Epstein, Gary Roberts, and Grace Beber. Springer, 2008, pp. 23–66.

[HK79] J. Michael Harrison and David M. Kreps. "Martingales and arbitrage in multiperiod securities markets". In: *Journal of Economic Theory* 20.3 (1979), pp. 381–408.

[Hau98] Espen Gaarder Haug. *The complete guide to option pricing formulas*. New York: McGraw-Hill, 1998.

[HTW19] Jakob Heiss, Josef Teichmann, and Hanna Wutte. *How implicit regularization of Neural Networks affects the learned function – Part I*. 2019. arXiv: `1911.02903 [cs.LG]`.

[Her17] Andrés Hernández. "Model Calibration with Neural Networks". In: *Risk* (2017).

[Hes93] Steven Heston. "A Closed-Form Solution for Options With Stochastic Volatility With Applications to Bond and Currency Options". In: 6 (1993), pp. 327–43.

[HRSW13] Norbert Hilber et al. *Computational methods for quantitative finance: finite element methods for derivative pricing*. Springer finance. Berlin: Springer, 2013.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[Hop82] John J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the National Academy of Sciences of the United States of America* 79 (8 1982), pp. 2554–2561.

[Hor91] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4.2 (1991), pp. 251–257.

[HSW90] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks". In: *Neural Networks* 3.5 (1990), pp. 551–560.

[HMT21] Blanka Horvath, Aitor Muguruza, and Mehdi Tomas. "Deep learning volatility: a deep neural network perspective on pricing and calibration in (rough) volatility models". In: *Quantitative Finance* 21.1 (2021), pp. 11–27.

[HNVW16] Tuomas Hytönen et al. *Analysis in Banach Spaces*. Vol. Volume I: Martingales and Littlewood-Paley Theory. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. Cham: Springer, 2016.

[Ito44] Kiyosi Ito. "Stochastic integral". In: *Proceedings of the Imperial Academy* 20.8 (1944), pp. 519–524.

[Jae10] Herbert Jaeger. *The "echo state" approach to analysing and training recurrent neural networks*. Tech. Rep. German National Research Center for Information Technology, 2010.

[JLL90] Patrick Jaillet, Damien Lamberton, and Bernard Lapeyre. "Variational inequalities and the pricing of American options". In: *Acta Applicandae Mathematica* 21 (3 1990), pp. 263–289.

[JLP20] Jasdeep Kalsi, Terry Lyons, and Imanol Perez Arribas. "Optimal Execution with Rough Path Signatures". In: *SIAM Journal on Financial Mathematics* 11.2 (2020), pp. 470–493.

[KR17] Ioannis Karatzas and Johannes Ruf. "Trading strategies generated by Lyapunov functions". In: *Finance and Stochastics* 21.3 (2017), pp. 753–787.

[KS98] Ioannis Karatzas and Steven Shreve. *Brownian Motion and Stochastic Calculus*. 2nd Corrected ed. 1998. Corr. 6th printing 2004. Berlin Heidelberg: Springer Science + Business Media, 1998.

[Kel75] John L. Kelley. *General topology*. Reprint. Vol. 27. Graduate texts in mathematics. New York: Springer-Verlag, 1975.

[KL20] Patrick Kidger and Terry Lyons. "Universal Approximation with Deep Narrow Networks". In: *Conference on Learning Theory 2020*. 2020.

[KB15] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, 2015, Conference Track Proceedings*. (San Diego, CA, USA). Ed. by Yoshua Bengio and Yann LeCun. 2015.

[KO19] Franz J. Kiraly and Harald Oberhauser. "Kernels for Sequentially Ordered Data". In: *Journal of Machine Learning Research* 20.31 (2019), pp. 1–45.

[KP98] Ron Kohavi and Foster Provost. "Glossary of terms". In: *Machine Learning* 30 (1998), pp. 271–274.

[KSV16] Yves-Laurent Kom Samo and Alexander Vervuurt. "Stochastic Portfolio Theory: A Machine Learning Perspective". In: *Capital Markets: Asset Pricing & Valuation eJournal* (2016).

[Kos10] Peter Kosmol. *Optimierung und Approximation*. 2., überarb. u. erw. Aufl. De Gruyter Studium. Berlin: De Gruyter, 2010.

[Kra96] Dmitry O. Kramkov. "Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets". In: *Probability Theory and Related Fields* 105 (4 1996), pp. 459–479.

[Kra21] Anastasis Kratsios. *Universal Regular Conditional Distributions*. 2021. arXiv: `2105.07743 [cs.LG]`.

[KB20] Anastasis Kratsios and Ievgen Bilokopytov. *Non-Euclidean Universal Approximation*. 2020. arXiv: `2006.02341 [cs.LG]`.

[KP21] Anastasis Kratsios and Leonie Papon. *Universal Approximation Theorems for Differentiable Geometric Deep Learning*. 2021. arXiv: `2101.05390 [cs.LG]`.

[KM97] Andreas Kriegl and Peter W. Michor. *The convenient setting of global analysis*. Vol. 53. Mathematical surveys and monographs. Providence, R.I: American Mathematical Society, 1997.

[Kry87] Nikolai Vladimirovich Krylov. *Nonlinear elliptic and parabolic equations of the second order*. Mathematics and its applications. Soviet series 7. Dordrecht: D. Reidel, 1987.

[LT05] S. Stig Larsson and Vidar Thomée. *Partial differential equations with numerical methods*. corr. 2nd print. Texts in applied mathematics 45. Berlin: Springer, 2005.

[Lla86] José G. Llavona. *Approximation of Continuously Differentiable Functions*. Vol. 130. North-Holland Mathematics Studies. North-Holland, 1986.

[LCL07] Terry J. Lyons, Michael Caruana, and Thierry Lévy. *Differential equations driven by rough paths*. Lecture notes in mathematics 1908. Berlin: Springer, 2007.

[LNP19] Terry Lyons, Sina Nejad, and Imanol Perez Arribas. "Numerical Method for Model-free Pricing of Exotic Derivatives in Discrete Time Using Rough Path Signatures". In: *Applied Mathematical Finance* 26.6 (2019), pp. 583–597.

[Mac77] Silvio Machado. "On Bishop's generalization of the Weierstrass-Stone theorem". In: *Indagationes Mathematicae* 80.3 (1977), pp. 218–224.

[Mar52] Harry Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[MP43] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5 (4 1943), pp. 115–133.

[Mer69] Robert C. Merton. "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case". In: *The Review of Economics and Statistics* 51.3 (1969), pp. 247–257.

[Mit97] Tom M. Mitchell. *Machine Learning*. McGraw-Hill series in computer science. Boston MA: WCB McGraw-Hill, 1997.

[Nac91] Leopoldo Nachbin. "On the Weighted Approximation of Continuously Differentiable Functions". In: *Proceedings of the American Mathematical Society* 111.2 (1991), pp. 481–485.

[Nac49] Leopoldo Nachbin. "Sur les algèbras denses de fonctions différentiables sur une variété". In: *Comptes rendus de l'Académie des Sciences de Paris* 228 (1949), pp. 1549–1551.

[Nac65] Leopoldo Nachbin. "Weighted Approximation for Algebras and Modules of Continuous Functions: Real and Self-Adjoint Complex Cases". In: *Annals of Mathematics* 81.2 (1965), pp. 289–302.

[Øks03] Bernt Øksendal. *Stochastic Differential Equations*. 6th. Berlin Heidelberg: Springer Science + Business Media, 2003.

[PW14] Sumit Pal and Ting-Kam Leonard Wong. "The geometry of relative arbitrage". In: *Mathematics and Financial Economics* 10 (2014), pp. 263–293.

[Pha09] Huyên. Pham. *Continuous-time Stochastic Control and Optimization with Financial Applications*. 1st ed. Stochastic Modelling and Applied Probability, 61. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[Pin99] Allan Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta Numerica* 8 (1999), pp. 143–195.

[Pro77] João B. Prolla. *Approximation of Vector Valued Functions*. North-Holland Mathematics Studies 25. Amsterdam: North-Holland, 1977.

[Pro71] João B. Prolla. "Bishop's generalized Stone-Weierstrass theorem for weighted spaces". In: *Mathematische Annalen* 191 (1971), pp. 283–289.

[PG76] João B. Prolla and Claudia S. Guerreiro. "An extension of Nachbin's theorem to differentiable functions on Banach spaces with the approximation property". In: *Arkiv för Matematik* 14.1-2 (1976), pp. 251–258.

[Pro05] Philip Protter. *Stochastic integration and differential equations*. 2nd ed., 3rd print. Vol. 21. Stochastic modelling and applied probability. Berlin: Springer, 2005.

[RR07] Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines". In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*. NIPS'07. Vancouver, British Columbia, Canada: Curran Associates Inc., 2007, pp. 1177–1184.

[RS05] James O. Ramsay and Bernard W. Silverman. *Functional data analysis*. 2nd ed. Springer series in statistics. New York: Springer, 2005.

[Ran84] Thomas J. Ransford. "A short elementary proof of the Bishop-Stone-Weierstrass theorem". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 96.2 (1984), pp. 309–311.

[Rau91] Jeffrey Rauch. *Partial Differential Equations*. 1st ed. 1991. Graduate Texts in Mathematics, 128. New York, NY: Springer New York, 1991.

[RS06] Michael Röckner and Zeev Sobol. "Kolmogorov equations in infinite dimensions: Well-posedness and regularity of solutions, with applications to stochastic generalized Burgers equations". In: *Annals of Probability* 34.2 (2006), pp. 663–727.

[Rud91] Walter Rudin. *Functional analysis*. 2nd ed. International series in pure and applied mathematics. Boston, Mass: McGraw-Hill, 1991.

[RW19] Johannes Ruf and Weiguan Wang. "Neural networks for option pricing and hedging: a literature review". In: *Journal of Computational Finance* 24.1 (2019), pp. 1–46.

[RX18] Johannes Ruf and Kangjianan Xie. "Generalised Lyapunov Functions and Functionally Generated Trading Strategies". In: *arXiv e-prints* (2018). arXiv: `1801.07817`.

[RN16] Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Third edition, global edition. Prentice Hall series in artificial intelligence. Boston: Pearson, 2016.

[Sam59] Arthur L. Samuel. *Some Studies in Machine Learning Using the Game of Checkers*. Tech. rep. 3. IBM Journal of Research and Development, 1959, pp. 210–229.

[SW99] Helmut H. Schaefer and Manfred P. Wolff. *Topological vector spaces*. 2nd ed. Vol. 3. Graduate texts in mathematics. New York: Springer, 1999.

[SZ06] Anton Maximilian Schäfer and Hans Georg Zimmermann. "Recurrent Neural Networks Are Universal Approximators". In: *Artificial Neural Networks – ICANN 2006*. Ed. by Stefanos D. Kollias et al. Berlin, Heidelberg: Springer, 2006, pp. 632–640.

[SSV18] Alexander Schied, Leo Speiser, and Iryna Voloshchenko. "Model-free portfolio theory and its functional master formula". In: *SIAM Journal on Financial Mathematics* 9 (2018), pp. 1074–1101.

[SV18] Alexander Schied and Iryna Voloshchenko. *The associativity rule in pathwise functional Ito calculus*. 2018. arXiv: `1605.08861 [math.PR]`.

[Sha64] William F. Sharpe. "Capital asset prices: a theory of market equilibrium under conditions of risk". In: *Journal of Finance* 19.3 (1964), pp. 425–442.

[Son06] Dieter Sondermann. "Introduction to Stochastic Calculus for Finance : A New Didactic Approach". In: *Lecture Notes in Economics and Mathematical Systems*. Vol. 579. Berlin: Springer-Verlag, 2006.

[Son05] Sneha Soni. "Applications of ANNs in Stock Market Prediction: A Survey". 2005.

[Sto48] Marshall Harvey Stone. "The Generalized Weierstrass Approximation Theorem". In: *Mathematics Magazine* 21.4 (1948), pp. 167–184.

[Str12] Winslow Strong. "Generalizations of Functionally Generated Portfolios with Applications to Statistical Arbitrage". In: *ArXiv e-prints* (2012). arXiv: `1212.1877`.

[Str14] Michael Struwe. *Funktionalanalysis I und II*. Lecture Notes in Fall Semester 2013 and Spring Semester 2014, ETH Zurich. 2014.

[SS18] Niladri Syam and Arun Sharma. "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice". In: *Industrial Marketing Management* 69 (2018), pp. 135–146.

[Tou13] Nizar Touzi. *Optimal Stochastic Control, Stochastic Target Problems, and Backward SDE*. 1st ed. Fields Institute Monographs, 29. New York: Springer, 2013.

[Tur50] Alan M. Turing. "Computing Machinery and Intelligence". In: *Mind* LIX.236 (1950), pp. 433–460.

[Wea99] Nik Weaver. *Lipschitz Algebras*. Second edition. Singapore: World Scientific, 1999.

[Wei85] Karl Weierstrass. "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen". In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin* II (1885).

[Whi88] Halbert White. "Economic prediction using neural networks: the case of IBM daily stock returns". In: *IEEE 1988 International Conference on Neural Networks*. Vol. 2. 1988, pp. 451–458.

[Wie32] Norbert Wiener. "Tauberian Theorems". In: *Annals of Mathematics* 33.1 (1932), pp. 1–100.

[Won15] Ting-Kam Leonard Wong. "Universal portfolios in stochastic portfolio theory". In: *ArXiv e-prints* (2015). arXiv: 1510.02808.

Philipp Schmocker
Curriculum vitae

Personal Information

Born on March 25th, 1994 in Berne
Citizenship: Switzerland; Place of citizenship: Beatenberg BE

Education

PH.D. IN ECONOMICS AND FINANCE — *Jun 2018 - present*
University of St.Gallen (HSG)
Dissertation: Universal Approximation on Path Spaces and Applications in Finance

M.SC. ETH IN MATHEMATICS — *Sep 2015 - Jun 2017*
Swiss Federal Institute of Technology Zurich (ETH)
Master Thesis: Frictions in Stochastic Portfolio Theory

B.SC. ETH IN MATHEMATICS — *Sep 2012 - Sep 2016*
Swiss Federal Institute of Technology Zurich (ETH)
Bachelor Thesis: Portfolio Optimization

MATURA — *Sep 2008 - Jun 2012*
Gymnasium Köniz-Lerbermatt, Köniz
Matura Thesis: Enigma – Construction and Background of a Cipher Machine

Employment

RESEARCH INTERNSHIP — *Feb 2020 - Aug 2020*
at Nanyang Technological University (NTU), Singapore

PH.D. STUDENT / ASSISTENT — *Jun 2018 - Jun 2021*
at University of St.Gallen (HSG)

INTERNSHIP MARKET INVESTIGATION — *Aug 2017 - May 2018*
at Swiss National Bank (SNB), Berne