Doctoral Thesis

Theory of Neural Networks with Applications in Finance, Insurance and Climate-Economy Modelling

Submitted in partial fulfillment of the requirements for the degrees of

Doctor of Natural Sciences
Doctor of Philosophy

Within the doctoral programs

Technical Mathematics
Actuarial Studies and Business Analytics

Written as part of a Cotutelle agreement between

TU Wien

Institute of Statistics and Mathematical Methods in Economics
Faculty of Mathematics and Geoinformation

Macquarie University

Department of Actuarial Studies and Business Analytics
Macquarie Business School

Under the supervision of

Prof. Dr. Uwe Schmock
Prof. Dr. Pavel V. Shevchenko

By

Dipl.-Ing. Aleksandar Arandjelović

Matriculation number 01425305
Student ID 46675868

Vienna, January 2025

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Abstract

This thesis sits at the intersection of three key research areas: (1) stochastic calculus, (2) artificial intelligence, with an emphasis on deep learning, and (3) financial and actuarial mathematics. In recent years, machine learning has increasingly influenced quantitative finance and actuarial science, with fascinating insights and applications. Topics such as the hedging of risky positions, synthetic scenario generation, and model calibration have all benefited from machine learning, making it possible to study previously intractable problems with remarkable precision and computational efficiency. The goal of this thesis is to advance the existing literature by introducing new results in the theory of neural networks from the point of view of financial and actuarial mathematics, and to exemplify these contributions through interesting case studies, which can be broadly grouped into the following three main thematic areas.

The first theme, *Deep Measure Projections*, focuses on the optimal projection of a given measure onto a set of algorithmically generated measures, a concept that aligns well with the existing quantitative finance literature. Incomplete financial markets typically require the choice of an appropriate pricing measure for arbitrage-free pricing of financial derivatives. Similarly, variance reduction techniques for Monte Carlo (MC) methods seek efficient sampling measures for computing MC estimators. Chapter 1 addresses this by studying variance reduction through changes in sampling measures that are calculated via feedforward neural networks.

The second theme, *Deep Surrogate Models*, examines the use of deep neural networks to approximate complex input-output maps, a method that is theoretically justified by universal approximation theorems. This approach is particularly useful when solving high-dimensional stochastic control problems, often encountered in quantitative finance and actuarial science, by least-squares Monte Carlo (LSMC) methods. Chapter 2 studies optimal insurance purchases when bequest motives are age-varying and life insurance and life annuities both carry loads, revealing up to two distinct periods of non-participation. Chapter 3 then extends the application of the LSMC method to complex stochastic climate-economy models, demonstrating how deep neural networks can improve the accuracy and efficiency of optimal policy

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

derivation in uncertain, high-dimensional environments.

The final theme, *Algorithmic Strategies*, involves the use of neural networks to approximate optimal decisions, such as those faced in dynamic trading, hedging, or reinsurance strategies, in feedback form. This approach, particularly deep hedging, has become a cornerstone method in quantitative finance and actuarial science for generating computationally feasible algorithms that identify optimal strategies. Chapter 4 introduces an application to algorithmic reinsurance policies that optimize the expected utility of terminal wealth perturbed by a modified Gerber–Shiu penalty function. Finally, Chapter 5 establishes universal approximation theorems for algorithmically generated stochastic (integral) processes, demonstrating that a stochastic calculus can be developed using algorithmic strategies.

Keywords: Deep learning; Universal approximation; Measure projection; Importance sampling; Surrogate model; Deep annuity puzzle; Least-squares Monte Carlo; Stochastic dynamic integrated climate-economy (DICE) model; Algorithmic strategy; Reinsurance; Stochastic integration; No free lunch with vanishing risk

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Acknowledgements

First and foremost, I extend my heartfelt gratitude to Pavel V. Shevchenko, my primary academic advisor, for his invaluable guidance, support, and encouragement throughout my academic journey over the past four years. His mentorship has opened many doors, leading to fruitful collaborations and opportunities, including the memorable experience of conducting research in Sydney.

I would like to thank Uwe Schmock, whose mentorship has been a constant and invaluable source of insight, beginning with my diploma thesis and continuing to this day through our collaborative work, which has always been a rewarding learning experience to me.

My sincere thanks go to Sandra Trenovatz, whose unwavering support smoothed the process of writing this dissertation, traveling to conferences, and fully engaging with my studies. Her skill in resolving any administrative challenge made my journey significantly easier.

I am grateful for the inspiring collaborations I have enjoyed over the years, particularly with Tomoko Matsui, Julia Eisenberg, Thorsten Rheinländer, Geoff Kingston as well as Gonçalo dos Reis. I owe special thanks to Geoff, who contributed his expertise to Chapter 2, and to Pavel and Julia, who prepared drafts of the introductions to Chapters 3 and 4. Uwe provided valuable feedback on a draft of Chapter 5, significantly enhancing the quality of the results. Each chapter is based on research papers I have authored or co-authored: Chapter 1 has been published in Finance and Stochastics (Arandjelović et al., 2025), and Chapter 2 has been published in the Journal of Economic Dynamics and Control (Arandjelović et al., 2023); Chapters 3 and 4 are under review, and Chapter 5 is being prepared for submission.

I gratefully acknowledge the financial support provided through the International Cotutelle Macquarie University Research Excellence Scholarship, as well as funding from the Institute of Statistical Mathematics in Japan.

Lastly, I extend my deepest gratitude to my family for their unwavering support, standing by my side through challenging times and celebrating my achievements. I would not be where I am today without you, and for this, I am truly grateful.

Aleksandar Arandjelović

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table of Contents

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

List of Figures

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

List of Tables

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Background and Motivation

In 1676, Gottfried Wilhelm Leibniz introduced the chain rule for derivatives of compositions of differentiable functions (Leibniz, 1676). More than one and a half centuries later, Augustin–Louis Cauchy suggested an early form of the gradient descent method (Cauchy, 1847). Fast-forward to the 1940s and 1950s, Norbert Wiener laid the groundwork for self-regulating mechanisms (Wiener, 1948), while Vladimir I. Arnold and Andrey N. Kolmogorov demonstrated that any continuous multivariate function can be expressed as a finite composition of continuous univariate functions and addition (Kolmogorov, 1956; Arnold, 1957; Kolmogorov, 1957), thereby solving a version of David Hilbert's thirteenth problem. By the late 1980s and early 1990s, George Cybenko and Kurt Hornik solidified the universal approximation property of feedforward neural networks (Cybenko, 1989; Hornik, 1991), and Sepp Hochreiter introduced together with Jürgen Schmidhuber long short-term memory (LSTM) networks soon thereafter (Hochreiter and Schmidhuber, 1997). It is without a doubt that, building on hundreds of years of dedicated research, artificial intelligence (AI), and in particular deep learning, has evolved into a powerful and pervasive tool in modern life.

The development of deep neural networks (DNNs) can be traced to the 1940s, when McCulloch and Pitts (1943) proposed the perceptron, a model for linear classification using a weighted sum of inputs and a step activation function. The 1970s and 1980s saw a major breakthrough with the development of the backpropagation algorithm (Linnainmaa, 1970; Werbos, 1974; Rumerlhart et al., 1986), which allowed for efficient training of multilayer networks through gradient descent and the chain rule. This period also saw the mathematical formalization of the universal approximation theorem, demonstrating that feedforward neural networks with a single hidden layer can approximate any continuous function. Despite periods of decline in interest in AI during the so-called "AI winters", advances in computational power, particularly with GPUs, and large-scale datasets have led to the resurgence and prominence of neural networks in recent years.

Given its computational strength, it was inevitable that deep learning would make its way into finance and insurance. In finance, it is applied to tasks like predicting stock market trends, credit risk assessment, and portfolio optimization. In insurance, it enhances underwriting, risk evaluation, and fraud detection. Additionally, generative models such as generative adversarial networks (GANs) are increasingly used to create synthetic data for stress testing and other risk management purposes. Integrating deep learning with traditional mathematical tools such as stochastic calculus allows for deeper insights into neural networks within the context of financial and actuarial mathematics.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

From a functional-analytic perspective, deep neural networks represent a simple class of functions that can be trained efficiently using modern numerical methods. Stochastic calculus, on the other hand, forms the backbone of much of quantitative finance and actuarial science. This naturally motivates the study of neural networks from the point of view of stochastic calculus, with the goal of deriving interesting case studies in the domain of quantitative finance and insurance. This thesis aims to contribute to the literature by introducing new theoretical results on neural networks in financial and actuarial mathematics, and demonstrating these findings through interesting case studies. We aim to contribute to three main thematic areas, which are described in the following sections.

Deep measure projections

The concept of deep measure projections refers to the task of optimally projecting a given probability measure onto a set of algorithmically generated measures using deep learning techniques. The "projection" process typically involves training a neural network to learn a mapping from a base probability measure to an optimal measure by minimizing a relevant loss function related to pricing errors or other criteria, such as expected utility or variance. Typically, the problem amounts to approximating the optimal density process, which describes how to adjust probabilities between the base and target measure. These approximations can be difficult to determine analytically or computationally via traditional methods, especially in high-dimensional settings with complex dynamics. Deep learning provides more efficient optimizations of these projections, offering solutions to problems such as variance reduction in Monte Carlo simulations, pricing, and model calibration.

In incomplete financial markets, the task of selecting an appropriate pricing measure is crucial for arbitrage-free pricing, as infinitely many equivalent martingale measures typically exist. Unlike in complete markets, where a unique risk-neutral measure defines fair pricing, in incomplete markets, the optimal measure must be chosen based on individual preferences while ensuring no arbitrage. Common approaches, such as the minimal-entropy martingale measure or the minimal martingale measure, often rely on minimizing a distance between the objective and risk-neutral probability measures. However, finding these optimal measures can be computationally challenging, making deep learning methods an attractive solution for such tasks.

Importance sampling, a variance reduction technique, is used to improve the efficiency of Monte Carlo simulations by altering the sampling measure. In option pricing, the goal is to sample those paths more frequently from scenarios that have a larger impact on the payoff of the option, while maintaining unbiased estimates. The choice of optimal sampling measures is particularly crucial for pricing out-of-the-money options, tail risk estimation, and other rare-event-driven instruments, where traditional Monte Carlo

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

methods struggle to capture low-probability but high-impact events. This technique is frequently applied in pricing exotic options and modelling tail risk in large insurance portfolios.

Chapter 1 extends the notion of deep measure projections to the importance sampling problem for pricing complex, high-dimensional and path-dependent European options. We model price processes X via stochastic differential equations (SDEs) of the form $\mathrm{d}X_t = a_t(X)\,\mathrm{d}[M]_t + b_t(X)\,\mathrm{d}M_t$, where M is a continuous local martingale with deterministic covariation $[M]$. The covariation process defines a model-dependent Hilbert space Λ^2, extending the space of square-integrable functions to the vector-valued case. Upon proving a universal approximation theorem in this generalized space, we construct algorithmically generated functions that are dense in the Cameron–Martin space H of the Gaussian measure that is induced by M on path space.

Once the universal approximation property in Cameron–Martin space has been established, we focus on pricing path-dependent claims $F(X)$ via Monte Carlo simulations. We induce a change of the sampling measure using density processes represented by stochastic exponentials $\mathcal{E}(f \bullet M)$, where neural networks f are integrated with respect to M. By Girsanov's theorem, this measure change corresponds to adding a drift to M that belongs to the algorithmically generated dense subset of H. By restricting the optimization to drifts from H, we prove that algorithmically generated drifts can approximate the optimal variance-minimizing drift arbitrarily well. These theoretical results are supported by extensive numerical simulations in pricing path-dependent European options under models that incorporate factors such as changing business activity, dynamic correlations, knock-out barriers, and high-dimensional asset baskets.

Deep surrogate models

Deep surrogate models refer to the use of deep neural networks to approximate complex input-output maps that arise in high-dimensional problems, such as pricing financial derivatives or solving stochastic control problems. These models act as computationally efficient substitutes for expensive numerical methods like Monte Carlo simulations or partial differential equation solvers, which are often infeasible in real-time applications. In essence, the surrogate model is trained on a large dataset of precomputed results and, once trained, can be deployed to provide near-instantaneous predictions or policy decisions, which is critical in real-time applications like hedging or portfolio management. Additionally, deep surrogate models can be integrated into the least-squares Monte Carlo (LSMC) method to approximate value functions and optimal control policies. This is particularly useful in high-dimensional problems like dynamic asset allocation or climate-economy models, where standard regression techniques may suffer from the curse of dimensionality and thus be inefficient or intractable.

When solving stochastic control problems via dynamic programming, one of the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

main challenges arises from the recursive Bellman equation, which requires computing conditional expectations of the value of a future state given the current state. In the context of pricing American options, Longstaff and Schwartz (2001) addressed this issue by proposing a least-squares Monte Carlo (LSMC) approach, where the conditional expectation – referred to as the continuation value – represents the value of the option if it is not exercised. The continuation value is approximated using polynomial regressions based on simulated paths, which works well in low-dimensional settings where the state is exogenous, i.e. not influenced by the control policy. However, this approach becomes problematic as the dimension of the state space increases or when the control policy itself influences the states, as the selection of appropriate basis functions becomes increasingly difficult and subjective. Since the conditional expectation can be viewed as an orthogonal projection in the Hilbert space of square-integrable random variables, neural networks, due to their universal approximation properties, offer a robust alternative to classical basis functions for high-dimensional, path-dependent, or complex control problems.

As a preliminary exercise in the field of optimal stochastic control, Chapter 2 explores a life-cycle model of optimal insurance purchases. In practice, very few retirees choose to annuitize a substantial portion of their wealth, even though rational choice models would predict them to do so – a phenomenon commonly known as the "annuity puzzle". By introducing load factors on both life insurance and life annuities alongside age-varying bequest motives, we demonstrate numerically that a thin market for annuities can coexist with a thick market for life insurance. More precisely, our simulations reveal up to two distinct periods of non-participation: one in midlife and the other adjoining the maximum age. Since loads on life-contingent insurance products can be interpreted as a form of illiquidity or transaction cost, the presence of non-participation periods aligns with existing literature on investment in risky assets under transaction costs, where similar non-transaction regions were identified.

Chapter 3 extends the LSMC method to stochastic climate-economy models, which integrate the stochastic evolution of climate variables like temperature and emissions with economic factors such as output, consumption and investment. These models are typically used in policy analysis, guiding decisions on carbon pricing, emissions regulation, and adaptation strategies to manage climate risks. However, their high dimensionality and complex variable interactions make them computationally challenging, as they are often formulated as stochastic dynamic programming problems. To overcome these challenges, we replace standard regression techniques in the LSMC method with deep neural network approximations, and demonstrate that the deep LSMC approach efficiently derives optimal policies for climate-economy models in the presence of uncertainty.

Solving stochastic climate-economy models offers significant potential for quantitative finance and actuarial science, particularly in the context of climate stress testing. In finance, climate stress testing is increasingly important for evaluating the resilience of portfolios through all channels of credit, market and liquidity risks, and pricing climate-related risks. For actuarial science, these models can help assess the potential impact

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of climate events on insurance liabilities and mortality rates, guiding product design and reserving. However, current methods for climate stress testing are inadequate, often relying on deterministic scenarios that fail to capture the stochastic nature of both climate and economic interactions. The results obtained in Chapter 3 therefore contribute to resolving a clear research gap: the need for computational methods capable of handling stochastic climate-economy models, and we leave the exploration of potential applications to climate stress testing to future work.

Algorithmic strategies

Algorithmic trading refers to the use of algorithms to automatically execute trades based on predefined rules and criteria. By automating the decision-making process, it aims to minimize human intervention, reducing emotional biases and increasing the speed, accuracy, and efficiency of trade executions. A prominent class of algorithmic strategies, known as "statistical arbitrage", exploits price inefficiencies or temporary imbalances between related financial instruments to generate profit. Another widely used strategy, "market making", continuously quotes buy and sell prices, capturing the bid-ask spreads, and providing liquidity to the market. However, poorly designed algorithms can cause unintended market impact, such as excessive market volatility or even flash crashes which – due to the widespread adaptation of algorithmic strategies – raises concerns about market stability and systemic risk.

Buehler et al. (2019) provide a rigorous mathematical framework for the optimality of algorithmic hedging strategies in a discrete-time setting, grounded in a simple yet profound application of the Doob–Dynkin representation lemma. This lemma allows strategies to be expressed as functions of the process generating the market information, i.e., the filtration with respect to which all processes are assumed to be adapted. Neural networks offer a natural tool for approximating these functions, with universal approximation theorems in L^p-spaces over finite Borel measure spaces justifying their use in this context. While other function classes could be used for these approximations, neural networks are particularly efficient due to their optimization via stochastic approximation techniques like stochastic gradient descent. Numerous simulation studies have demonstrated the efficiency with which algorithmic strategies can find optimal trading and hedging decisions.

Chapter 4 extends the concept of deep hedging to the reinsurance domain. Optimal reinsurance strategies are an active area of research with numerous publications studying different model specifications and preference structures. We propose a general framework that uses deep learning techniques to optimize reinsurance strategies by maximizing a target functional comprising a utility function penalized by an extended Gerber–Shiu function. This method allows the insurer to balance between maximizing the expected utility of terminal wealth, and minimizing the probability of ruin. The framework includes many classical models as special cases, and the theoretical results are illustrated by a

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

numerical simulation, where the surplus process is given by a Cramér–Lundberg model perturbed by a mean-reverting Ornstein–Uhlenbeck process.

Finally, Chapter 5 extends the concept of deep hedging from discrete- to continuous-time settings. While discrete-time models are often computationally simpler, they can suffer from discretization bias and may overlook critical intra-period dynamics, especially in high-frequency trading. Continuous-time models offer a more refined and mathematically elegant framework for option pricing and risk management, often yielding closed-form solutions that are difficult to obtain in discrete time. Another significant benefit is the ability to model jumps and sudden market movements via jump-diffusion or Lévy processes, capturing more realistic market dynamics that are difficult to represent in discrete-time frameworks. In short, studying algorithmic strategies in continuous time allows for the application of advanced tools from continuous-time stochastic calculus, which may potentially widen the scope of applicability of these strategies and motivate new solution approaches to old problems.

In the same way that measure theory defines the integral of general functions through limits of step functions, stochastic integrals of integrable processes are defined as limits over integrals of simple processes. By proving that a large class of stochastic processes, including simple processes, can be approximated by algorithmic strategies, it follows that stochastic integral processes can also be approximated by integrals of algorithmic strategies. This result highlights the fundamental role artificial neural networks play in the theory of stochastic integration, and yields several interesting implications for mathematical finance. Specifically, integrals of algorithmic strategies with respect to square-integrable martingale price processes are topologically dense in the stable subspace generated by these price processes, giving rise to the notion of deep mean-variance hedging in continuous time. Moreover, a sufficient condition for the existence of an equivalent martingale measure in the context of Stricker (1990, Theorem 3) can be linked to a no free lunch with vanishing risk condition (NFLVR) for algorithmic strategies, thereby allowing – for the first time – to discuss algorithmic strategies from a no-arbitrage perspective.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Part I.

Deep Measure Projections

1. Importance Sampling for Option Pricing with Feedforward Neural Networks

Monte Carlo methods are amongst the most essential tools for the numerical evaluation of financial derivatives. Classical asset pricing theory often calls for the computation of expectations of the form

$$\mathbb{E}_{\mathbb{P}}[F(X)] = \int_{\Omega} F(X(\omega))\,\mathbb{P}(\mathrm{d}\omega),$$

where F is a payoff functional, X is an asset price process solving a stochastic differential equation (SDE) of the form $\mathrm{d}X_t = a(X)\,\mathrm{d}C_t + b(X)\,\mathrm{d}M_t$ for $t \in [0,T]$ with a finite time horizon $T > 0$ for some potentially path-dependent coefficients a, b, a continuous process C of locally finite variation, and a continuous local martingale M, and $\mathbb{P}$ is a probability measure on a measurable space $(\Omega, \mathcal{F})$. By averaging the payoffs over randomly sampled trajectories of X, one can estimate the price in many cases where no analytic solution for $\mathbb{E}_{\mathbb{P}}[F(X)]$ is available.

The variance of the Monte Carlo estimator is inversely proportional to the number of trajectories simulated and proportional to the variance of the option payoff. The square root of this variance is referred to as the standard error and in principle, it can be made as small as needed by simulating a sufficiently high number of trajectories. However, given limitations on computational time, the error can still be too large to be acceptable, especially for further calculations of option price derivatives (the so-called Greeks) required for hedging and risk management.

The usage of variance reduction methods can drastically reduce this error. There are many different methods to reduce the variance of Monte Carlo estimators, one of which is importance sampling. This method is based on changing the sampling measure from which the trajectories are generated from $\mathbb{P}$ to some equivalent measure $\mathbb{P}_h$, thereby overweighting important scenarios to increase the numerical efficiency of the estimates. Due to Girsanov's theorem, this corresponds to adding a drift $h \in H$ to the process M, where H denotes a prescribed space of functions or processes from which the drift adjustment is chosen. One then writes

$$\mathbb{E}_{\mathbb{P}}[F(X)] = \mathbb{E}_{\mathbb{P}_h}[F(X)Z_h^{-1}],$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

where Z_h denotes a Radon–Nikodým density of $\mathbb{P}_h$ with respect to $\mathbb{P}$, which also depends on the time-horizon T. Instead of simulating realizations of $F(X)$ with respect to $\mathbb{P}$, one then simulates realizations of $F(X)Z_h^{-1}$ with respect to $\mathbb{P}_h$, and chooses $h \in H$ such that the variance of $F(X)Z_h^{-1}$ is minimized.

While this method usually requires a lot of specific knowledge about the model at hand, it has the potential to drastically reduce the variance of the corresponding Monte Carlo estimator. In other words, importance sampling is a powerful method that involves the complex optimization problem of choosing an appropriate sampling measure which minimizes the variance of the Monte Carlo estimators.

Neural networks provide an algorithmically generated class of functions which on the one hand enjoy the universal approximation property in many different topological spaces, meaning that they are dense in these spaces, and on the other hand can be trained in a numerically efficient way. Having recently entered the realm of mathematical finance, neural networks are successfully used e.g. for model calibration, hedging and pricing. This chapter develops a method that uses feedforward neural networks to perform importance sampling for complex stochastic models, which applies in particular to the evaluation of path-dependent derivatives. By optimizing over drifts from a dense subspace $H(D)$ of H that is generated by a set D of feedforward neural networks, we obtain a tractable problem which is both theoretically justified and numerically efficient.

In Section 1.1, we characterize tractable spaces H from which the drift adjustments may be chosen, and study their analytic properties. Whenever M is a vector-valued continuous local martingale with deterministic covariation, it induces a Gaussian measure, to which one can assign a Hilbert space H, the Cameron–Martin space. Due to the multivariate nature of our study, we recall in Lemma 1.10 some concepts which originate from the theory of stochastic integration with respect to vector-valued semimartingales. A detailed characterization of the corresponding Cameron–Martin space H that is induced by M is provided in Proposition 1.18, where the general formulation allows us to specifically incorporate complex and time-inhomogeneous covariance patterns for M into our models. Theorem 1.24 then yields the essential approximation result that characterizes dense subspaces $H(D)$ of H which are generated by prescribed sets of functions D in an abstract and general setting, and in particular applies to sets D of feedforward neural networks as a special case.

In Section 1.2, we focus our attention on feedforward neural networks, where we distinguish between neural networks of deep, narrow and shallow kind. Propositions 1.28 and 1.29 yield two approximation results which provide a theoretical justification for considering sets D that consist of feedforward neural networks. Example 1.31 then shows in a classical setting that the set $H(D)$ which is generated by a set D of feedforward neural networks has an explicit and tractable characterization. As a direct consequence of Theorem 1.24, Subsection 1.2.1 then discusses a result which in particular implies that every smooth function can, up to an isometry, be approximated by feedforward neural networks arbitrarily well with respect to Hölder type topologies, which are stronger than

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

the topology of uniform convergence.

Section 1.3 contains a detailed study of the importance sampling problem, where we aim to minimize the variance of $F(X)Z_h^{-1}$ with respect to $\mathbb{P}_h$ by approximating the optimal drift h with a feedforward neural network. Theorem 1.38 proves that the functional $V\colon H \to \mathbb{R}_+$ which needs to be minimized is, under suitable generic assumptions, continuous and admitting a minimizer $h^* \in H$, which can be approximated, up to an isometry, arbitrarily well by feedforward neural networks. Here, we not only prove convergence to the optimal drift h^*, but we moreover show that the corresponding Radon–Nikodým densities converge to Z_{h^*}. To this end, we prove that feedforward neural networks induce equivalent probability measures, whose densities with respect to the original measure converge in L^p-spaces, see Lemma 1.37. Moreover, Subsection 1.3.1 contains a discussion of a classical importance sampling approach that utilizes results from the theory of large deviations, where we show that feedforward neural networks can be employed to solve the corresponding variational problem which appears in this approach. Let us note that, while the results from Sections 1.1 and 1.2 are applied to importance sampling in Section 1.3, they are also of independent interest.

Section 1.4 contains a comprehensive numerical study pricing path-dependent European options for asset price models that incorporate factors such as changing business activity, knock-out barriers, dynamic correlations, and high-dimensional baskets. Section 1.5 contains a brief glimpse at the theory of Gaussian measures and collects the proofs of all results.

The line of research which eventually lead up to the present work originates from Glasserman et al. (1999). The authors study the problem of pricing path dependent options by using techniques from the theory of large deviations to perform a change of sampling measure that reduces the standard error of the Monte Carlo estimator. Moreover, the authors use stratified sampling in order to further improve their simulations, and while we will not be using this technique, the interested reader might try to add stratified sampling on top of the method which we outline below.

The main motivation for this work was provided by Guasoni and Robertson (2008). As in Glasserman et al. (1999), the authors employ methods from the theory of large deviations to obtain a variational problem whose solution yields an asymptotically optimal drift adjustment. The main difference to the present work is that we do not pass to a small noise limit. However, as it turns out, our method also complements the method presented in Guasoni and Robertson (2008), see Subsection 1.3.1 for further details.

An extension of the methods used in Guasoni and Robertson (2008) to the study of importance sampling for stochastic volatility models has been provided in Robertson (2010). Note that our method does apply to these types of models as well, see Example 1.34 in Section 1.3 and Section 1.4, where we provide simulation results for several stochastic volatility models.

Another interesting contribution is dos Reis et al. (2023), where the authors study importance sampling for McKean–Vlasov SDEs. Similarly as in Guasoni and Robertson

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

(2008); Robertson (2010) methods from the theory of large deviations yield an asymptotically optimal drift adjustment, and the authors discuss two different methods for the simulation of the solution to the McKean–Vlasov SDE under a change of measure.

The idea to use methods from the theory of stochastic approximation for the purpose of importance sampling has been studied extensively in Lemaire and Pagès (2010). This paper heavily influenced Section 1.3, especially the proof of Theorem 1.38 relies partially on a straightforward extension of the proof of Lemaire and Pagès (2010, Proposition 4). Let us also note that, while the setting of Lemaire and Pagès (2010, Section 3) could be extended to our setting below, it might be of particular interest to understand how the algorithm proposed in Lemaire and Pagès (2010, Theorem 4) could be adapted to the setting of Section 1.3 below in order to yield convergence of the stochastic gradient descent algorithm when training feedforward neural networks.

Finally, one very original contribution that studies measure changes which are induced by neural networks for the purpose of Monte Carlo simulations is Müller et al. (2019). The authors also study importance sampling and apply their results to light-transport simulations. The main difference to Müller et al. (2019) is that our method applies to the pricing of financial derivatives in a mathematically more natural way by using methods from the theory of stochastic calculus. Here, we focus on neural networks that are of feedforward type. For more details on the studied neural networks architectures and related literature, see Section 1.2.

Notation for this chapter

Unless stated otherwise, we endow $\mathbb{R}^d$ for each $d \in \mathbb{N}$ with the corresponding Euclidean norm $|\cdot|$. I_d denotes the identity matrix in $\mathbb{R}^{d\times d}$, and we write $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$. Given two vectors x, y of the same dimension, we denote by $x \odot y$ their Hadamard product. If Σ is a matrix, we denote by $\Sigma^\intercal$ its transpose. For $x \in \mathbb{R}^d_+$ and $p > 0$, we understand x^p to hold componentwise, and write $\sqrt{x}$ if $p = 1/2$. Let us also convene that $\inf\{\varnothing\} = \infty$. For each linear operator A between normed spaces, we denote by $\|A\|_{\mathrm{op}}$ its operator norm. If E_1, E_2 denote two metric spaces and D is a subset of E_1, we say that D is dense in E_2 up to an isometry, if there exists an isometry $J\colon E_1 \to E_2$, such that $J(D)$ is dense in E_2. If H denotes a Hilbert space, the notation $H^* \cong H$ is to indicate that we identify H^* with H via the isometric isomorphism given by the Fréchet–Riesz representation theorem.

Given a topological space $(S, \mathcal{T})$, we denote by S^* and S' the topological and algebraic dual spaces, respectively, and write $(f, x) = f(x)$ for $(f, x) \in S^* \times S$ as well as $\mathcal{B}_S$ for the Borel σ-algebra on S. Given $F \in \mathcal{T}$, we denote by F° and $\bar{F}$ the interior and closure of F, respectively. Whenever ν denotes a Borel measure on S, we say that $f\colon S \to \mathbb{R}^d$ is locally ν-essentially bounded, if $(\nu\text{-})\operatorname{ess\,sup}_{x\in K} |f(x)| < \infty$ for each compact $K \subset S$. Given an interval $[0, T]$, the space $C_0([0, T]; \mathbb{R}^d)$ consists of all $\mathbb{R}^d$-valued, continuous functions on $[0, T]$ that vanish at the origin.

Whenever $(S, \mathcal{S})$ is a measurable space, where S denotes a set and $\mathcal{S}$ denotes a σ-algebra

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

on S, we denote by $\mathcal{L}^0(\mathcal{S};\mathbb{R}^d)$ the space of $\mathbb{R}^d$-valued, $\mathcal{S}$-measurable functions on S. If μ is a measure on $\mathcal{S}$ and $f \in \mathcal{L}^0(\mathcal{S}) = \mathcal{L}^0(\mathcal{S};\mathbb{R})$, we denote by $f \cdot \mu$ the Lebesgue integral of f with respect to μ, provided that it exists. For $p > 0$, we further denote by $L^p(\mu)$ the space of equivalence classes of p-integrable functions from $\mathcal{L}^0(\mathcal{S})$. The law of a random variable Z is denoted by $\mathcal{L}(Z)$. If M is an $\mathbb{R}^d$-valued semimartingale, and $H \in L(M)$, we denote by $H^\intercal \bullet M$ the stochastic integral of H with respect to M. We denote by $\mathcal{N}(m,\Sigma)$ the normal distribution with expected value $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Finally, we denote for each $p \geq 1$ by $\mathcal{H}^p$ the Banach space of continuous L^p-integrable martingales, where the dependency on the underlying filtered probability space is implicit.

1.1. Universal approximation in Cameron–Martin space

In this section, we study a tractable space H whose elements will be used to adjust the drift of M for the purpose of importance sampling in Sections 1.3 and 1.4 below. Moreover, we identify dense linear subspaces of H and obtain an explicit characterization of the Cameron–Martin spaces of a large class of Gaussian measures, which is of independent interest. For details about Gaussian measures, we refer to Subsection 1.5.1.

Let $(\Omega,\mathcal{F},\mathbb{F},\mathbb{P})$ with $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ denote a filtered probability space, such that $\mathbb{F}$ contains all $\mathbb{P}$-null sets of $\mathcal{F}$. As index set for the time parameter, we consider $[0,T]$ with a finite time horizon $T > 0$. Without loss of generality, we may assume that $\mathcal{F}_T = \mathcal{F}$. We denote by λ the restriction of the Lebesgue–Borel measure to $[0,T]$, and fix two dimensions $d,n \in \mathbb{N}$. Let $M = (M_t)_{t\in[0,T]}$ be an $\mathbb{R}^d$-valued continuous local martingale with $M_0 = 0$. Unless stated otherwise, we assume all stochastic processes to be $\mathbb{F}$-adapted.

Let us start with a classical example that highlights the main concepts which are of importance in this section, before extending the study to a more general setting.

Example 1.1. Let (E,H,γ) denote the classical Wiener space, i.e. $E = C_0([0,T];\mathbb{R}^d)$, $H = \{h(t) = (\mathbb{1}_{[0,t]}f_h)\cdot\lambda, t\in[0,T] : f_h \in L^2(\lambda;\mathbb{R}^d)\}$ is the space of $\mathbb{R}^d$-valued, absolutely continuous functions on $[0,T]$ that admit a square-integrable density with respect to λ, and γ is the classical Wiener measure on E, which is the Borel probability measure on E that is induced by $\mathbb{R}^d$-valued standard Brownian motion $B = (B_t)_{t\in[0,T]}$.

Endowed with the inner product $\langle g,h\rangle_H = \langle f_g, f_h\rangle_{L^2(\lambda;\mathbb{R}^d)}$, one can show that $(H,\langle\cdot,\cdot\rangle_H)$ is a real separable Hilbert space, which is continuously embedded into E as a dense linear subspace. The operator $J\colon L^2(\lambda;\mathbb{R}^d) \to H, f_h \mapsto h(\cdot) = (\mathbb{1}_{[0,\cdot]}f_h)\cdot\lambda$ is a linear isometry by construction and thus continuous. Whenever D is a dense linear subspace of $L^2(\lambda;\mathbb{R}^d)$, it follows that $J(D)$ is a dense linear subspace of H and thus densely embedded into E. In other words, $\overline{J(D)} = H$ and $\overline{J(D)} = E$, where the closure of $J(D)$ is taken in H and E, respectively.

Section 1.1 is dedicated to a refined study of the identity $\overline{J(D)} = H$ in a generalized setting. To this end, let us state an assumption which allows us to study the process M

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

as a Gaussian process, and simplifies the proofs of Section 1.3. Moreover, it leads to a natural candidate for the space H of drift adjustments, which consists of deterministic functions (see Definition 1.13 below).

General Assumption 1.2. The covariation process $[M]$ is, up to indistinguishability, deterministic, and $\operatorname{tr}([M])_T > 0$ outside a $\mathbb{P}$-null set.

In what follows, we disregard the evanescent- and $\mathbb{P}$-null sets on which the two conditions from Assumption 1.2 are violated, and consider equalities between stochastic processes and (in)equalities between random variables to hold up to indistinguishability and $\mathbb{P}$-almost surely, respectively.

Definition 1.3. The quadratic variation process $C \coloneqq \operatorname{tr}([M])$, being increasing and of finite variation, induces a finite Lebesgue–Stieltjes measure on $([0,T], \mathcal{B}_{[0,T]})$, which we denote μ.

Remark 1.4. Due to Lévy's characterization (cf. Schmock (2024, Theorem 7.1)), the increments $M_t - M_s$ are independent of $\mathcal{F}_s$ with $\mathcal{L}(M_t - M_s) = \mathcal{N}(0, [M]_t - [M]_s)$ for all $s < t$ in $[0,T]$. Therefore, M is a centered Gaussian process, and Kallenberg (2021, Theorem 11.5) shows that M is an $\mathbb{F}$-Markov process. Moreover, M is a martingale, since $\mathbb{E}[M_t - M_s|\mathcal{F}_s] = \mathbb{E}[M_t - M_s] = 0$ for $s < t$ in $[0,T]$. Note that $\operatorname{Cov}(M_t, M_s) = [M]_{s\wedge t}$ for $s, t \in [0,T]$ since, assuming without loss of generality that $s < t$,

$$\operatorname{Cov}(M_t, M_s) = \mathbb{E}[M_t M_s^\intercal] = \mathbb{E}[(M_t - M_s)M_s^\intercal] + \operatorname{Cov}(M_s, M_s) = [M]_s.$$

Remark 1.5. Assumption 1.2 implies that $\mu([0,T]) > 0$.

We write $\mu = \mu_{\mathrm{a}} + \mu_{\mathrm{s}}$ for the Lebesgue decomposition of μ with respect to λ into an absolutely continuous measure $\mu_{\mathrm{a}} = f_\lambda \cdot \lambda$ and a singular measure μ_{s}, where f_λ denotes a Radon–Nikodým density of μ_{a} with respect to λ, and both μ_{a} and μ_{s} are finite measures. Note that μ has no atoms, since $[M]$ and therefore also C are continuous.

We now proceed in line with Cohen and Elliott (2015, Section 12.5). The covariation process $[M^i, M^j]$, being of finite variation, induces a finite signed (and due to the continuity of $[M^i, M^j]$ atomless) measure $\mu_{i,j}$ on $([0,T], \mathcal{B}_{[0,T]})$ for all $i, j \in \{1, 2, \dots, d\}$. It follows from the Kunita–Watanabe inequality for Lebesgue–Stieltjes integrals (cf. Schmock (2024, Lemma 5.91)), that the total variation measure $|\mu_{i,j}|$ is absolutely continuous with respect to μ, hence an application of the Radon–Nikodým theorem for signed measures (cf. Cohen and Elliott (2015, Section 1.7.14)) yields the existence of a real-valued density $\mathrm{d}\mu_{i,j}/\mathrm{d}\mu =: \pi_{i,j}$, that is in $L^1(\mu)$ since $\mu_{i,j}$ is finite.

We collect $(\pi_{i,j})_{i,j=1,\dots,d}$ into a measurable function π that assumes, due to the symmetry of $[M]$, values in the space of symmetric matrices in $\mathbb{R}^{d\times d}$, and write

$$[M]_t = (\pi \bullet C)_t = (\mathbb{1}_{[0,t]}\pi) \cdot \mu, \quad t \in [0,T], \tag{1.1}$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

where the notation is to be understood componentwise. Let $(\eta_k)_{k\in\mathbb{N}}$ be dense in $\mathbb{R}^d$. For each $k \in \mathbb{N}$, set $A_k = \{s \in [0,T] : \eta_k^\mathsf{T} \pi(s) \eta_k \geq 0\}$ and moreover $A = \bigcap_{k\in\mathbb{N}} A_k$. Note that $A = \{s \in [0,T] : \eta^\mathsf{T} \pi(s) \eta \geq 0, \forall \eta \in \mathbb{R}^d\}$ and

$$\int_0^t \eta_k^\mathsf{T} \pi(s) \eta_k \, \mu(\mathrm{d}s) = \left((\eta_k^\mathsf{T} \pi \eta_k) \bullet C\right)_t = [\eta_k^\mathsf{T} M]_t \geq 0, \quad k \in \mathbb{N}, \, t \in [0,T],$$

which implies that each A_k^c is a μ-null set, and therefore A^c, being the countable union of all sets A_k^c, is a μ-null set, too. We conclude that π is positive semi-definite μ-almost everywhere. Note that we could, in the spirit of Cherny and Shiryaev (2002); Jacod (1979, 1980); Memin (1980) and without loss of generality, replace π by $\tilde{\pi} = \pi \mathbb{1}_A$, and thus assume that it is positive semi-definite for each $t \in [0,T]$. For the purpose of this chapter, this step is not necessary though.

Remark 1.6. The decomposition of $[M]$ into a matrix-valued function π and an increasing process C is not unique. For example, take $\tilde{C} = \sum_{i=1}^d \eta_i [M^i]$, where $\eta \in \mathbb{R}^d$ is chosen such that $\eta_i > 0$ for all $i \in \{1, 2, \ldots, d\}$. More generally, take $\tilde{C} = \sum_{i=1}^d f_i \bullet [M^i]$, with $f_i \colon [0,T] \to \mathbb{R}_+ \setminus \{0\}$ in $L^1(\mu_{i,i})$ for each $i \in \{1, 2, \ldots, d\}$. In both cases, the corresponding function $\tilde{\pi}$ is then constructed as in Remark 1.5, and generally differs from π. Lemma 1.10(e) below will show that the non-uniqueness of (π, C) is not a problem though.

Example 1.7. If $\pi \equiv I_d$ and $\mu = \lambda$, then M is, by Lévy's characterization, an $\mathbb{R}^d$-valued standard Brownian motion.

Example 1.8. An example that holds relevance for practitioners is the multivariate Heston model, which we will briefly describe.

Let $d = 2n$ for some $n \in \mathbb{N}$. We consider a dynamic diffusion matrix given by $[0,T] \ni t \mapsto \Sigma(t) \in \mathbb{R}^{d\times d}$, a vector of appreciation rates $r \in \mathbb{R}^n$, an n-dimensional vector of positive mean-reversion levels m, and an $n \times n$-dimensional diagonal matrix with positive entries Θ representing mean-reversion speeds. To avoid degeneracy, we assume that for each $k \in \{1, 2, \ldots, n\}$ and $t \in [0,T]$, $(\Sigma_{k,\cdot}(t))^\mathsf{T}$ is not the zero vector. Let $M_t = \Sigma(t) B_t$, where B denotes a standard Brownian motion with values in $\mathbb{R}^d$. Note that $[M]_t = \mathrm{Cov}(M_t, M_t) = \int_0^t \Sigma(s) \Sigma^\mathsf{T}(s) \, \mathrm{d}s$ for each $t \in [0,T]$. Hence, in light of Remark 1.5 above, we may choose $\mu = \lambda$ and $\pi(t) = \Sigma(t)\Sigma^\mathsf{T}(t)$.

Fix two n-dimensional initial value vectors s, v with positive entries. For simplicity, we write $M^{(1)} = (M^1, M^2, \ldots, M^n)^\mathsf{T}$ as well as $M^{(2)} = (M^{n+1}, M^{n+2}, \ldots, M^{2n})^\mathsf{T}$ such that $M = (M^{(1)}, M^{(2)})^\mathsf{T}$. Let $X = (S, V)^\mathsf{T}$, and let the asset price follow the SDE $\mathrm{d}S_t = (r \odot S_t) \, \mathrm{d}t + (S_t \odot \sqrt{V_t}) \odot \mathrm{d}M_t^{(1)}$, subject to $S_0 \equiv s$. The n-dimensional instantaneous variance process V follows the Cox–Ingersoll–Ross (CIR) type SDE $\mathrm{d}V_t = \Theta(m - V_t) \, \mathrm{d}t + \sqrt{V_t} \odot \mathrm{d}M_t^{(2)}$, subject to $V_0 \equiv v$. Here, we see that asset price models whose dynamics are driven by multivariate Brownian motions with dynamic variance-covariance matrices

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

fall within the scope of our setting. More generally, one could think of replacing B_t by a time-changed Brownian motion $B_{f(t)}$ for a given deterministic time-change f.

In Section 1.4, we will study special cases of this model, where we will impose either a time-change to model changing levels of business activity, or a dynamic correlation structure.

Example 1.9. Set $T = 1$, and let B be a standard $\mathbb{R}^d$-valued $(\mathbb{G}, \mathbb{P})$-Brownian motion, where $\mathbb{G} = (\mathcal{G}_t)_{t\in[0,T]}$ denotes a filtration of $\mathcal{F}$. Let $f\colon [0,T] \to [0,T]$ be either Cantor's ternary function or Minkowski's question-mark function, and let $\Sigma \in \mathbb{R}^{d\times d}$ be a diffusion matrix. Recall that Cantor's ternary function is continuous, monotonically increasing, has derivative zero on a set of Lebesgue measure zero, but is not absolutely continuous. Likewise, Minkowski's question-mark function has the same properties, while being even strictly increasing. Set $M_t \coloneqq \Sigma B_{f(t)}$ for $t \in [0,T]$, and note that M is an $\mathbb{R}^d$-valued continuous $(\mathbb{F}, \mathbb{P})$-martingale with $M_0 = 0$ and $[M]_t = f(t)\Sigma\Sigma^\mathsf{T}$ for $t \in [0,T]$, where the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ is given by $\mathcal{F}_t = \mathcal{G}_{f(t)}$ for $t \in [0,T]$. The corresponding Lebesgue–Stieltjes measure μ is singular with respect to λ, and while $\operatorname{tr}([M])$ is increasing when f is Cantor's ternary function, $\operatorname{tr}([M])$ is even strictly increasing when f is Minkowski's question-mark function. See Salem (1943) for further examples of functions f that can be used for constructions of this kind.

Based on the pair (π, μ), we define a weighted L^2-space, which we denote Λ^2, which is a generalization of the space $L^2(\lambda; \mathbb{R}^d)$ in the context of Example 1.1, and recall some elementary properties. In Section 1.3, where we study importance sampling, functions f from Λ^2 will be used to construct equivalent measures via the Doléans exponential $\mathcal{E}(f^\mathsf{T} \bullet M)$. As we will argue in Section 1.2, feedforward neural networks are dense in Λ^2 under suitable assumptions which, due to Theorem 1.38, provides a theoretical justification for using feedforward neural networks in order to calibrate an optimal sampling measure that minimizes the variance of the Monte Carlo estimators in Sections 1.3 and 1.4. The definition of the space Λ^2 is inspired by the concept of vector stochastic integration, see Cherny and Shiryaev (2002); Jacod (1979); Memin (1980), and in particular Jacod (1979, Chapitre IV), for further details and generalizations.

Lemma 1.10. *Let Λ^2 denote the set of all $f \in \mathcal{L}^0(\mathcal{B}_{[0,T]}; \mathbb{R}^d)$ with*

$$\|f\|_{\Lambda^2} \coloneqq \Big(\int_0^T f^\mathsf{T}(s)\pi(s)f(s)\,\mu(\mathrm{d}s) \Big)^{1/2} < \infty,$$

where we identify $f, g \in \Lambda^2$ if $(f-g)^\mathsf{T}\pi(f-g) = 0$ μ-almost everywhere, and write $f \sim g$ in this case. We further set $\langle f, g\rangle_{\Lambda^2} \coloneqq \int_0^T f^\mathsf{T}(s)\pi(s)g(s)\,\mu(\mathrm{d}s)$ for $f, g \in \Lambda^2$. Then:

(a) $(\Lambda^2, \langle\cdot,\cdot\rangle_{\Lambda^2})$ is a real separable Hilbert space;

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

(b) To each $F \in (\Lambda^2)^*$ *there corresponds a unique function* $g \in \Lambda^2$*, such that*

$$F(f) = \int_0^T g^\mathsf{T}(s)\pi(s)f(s)\,\mu(\mathrm{d}s), \quad f \in \Lambda^2,$$

and $\|F\|_{\mathrm{op}} = \|g\|_{\Lambda^2}$*. Therefore,* $(\Lambda^2)^*$ *is isometrically isomorphic to* Λ^2*;*

(c) We denote by $\Lambda^{2,0}$ *the set of all* $f \in \mathcal{L}^0(\mathcal{B}_{[0,T]}; \mathbb{R}^d)$ *that satisfy* $f_i \in L^2(\mu_{i,i})$ *for each* $i \in \{1, 2, \ldots, d\}$*, where we identify functions in the same manner as above. Then:*

(1) $(\Lambda^{2,0}, \langle \cdot, \cdot \rangle_{\Lambda^2})$ *is a separable inner product space with* $\Lambda^{2,0} \subset \Lambda^2$*;*

(2) $\Lambda^{2,0}$ *is dense in* Λ^2*, hence* Λ^2 *is the completion of* $\Lambda^{2,0}$ *with respect to* $\|\cdot\|_{\Lambda^2}$*;*

(d) $(C([0,T]; \mathbb{R}^d), \|\cdot\|_\infty)$ *is continuously embedded into* $(\Lambda^{2,0}, \|\cdot\|_{\Lambda^2})$ *as a dense linear subspace, where* $\|f\|_\infty := \sup_{t\in[0,T]} |f(t)|$ *for* $f \in C([0,T]; \mathbb{R}^d)$*;*

(e) Λ^2 *and* $\Lambda^{2,0}$ *do not depend on the specific choice of* (π, μ) *that satisfy* (1.1)*.*

Example 1.11. According to Jacod (1979, Lemme 4.30) and the discussion thereafter, a sufficient condition for $\Lambda^{2,0} = \Lambda^2$ to hold is if there exists a constant $c > 0$ such that for all $f \in \Lambda^2$, $\sum_{i=1}^d \pi_{i,i} f_i^2 \le c f^\mathsf{T} \pi f$ holds μ-almost everywhere. Examples where this applies are when π is a diagonal or uniformly strictly elliptic matrix, where the latter condition means that there exists a constant $c > 0$ such that $c|\eta|^2 \le \eta^\mathsf{T}\pi\eta$ holds for all $\eta \in \mathbb{R}^d$.

Example 1.12. Let us state one example, which is a deterministic version of Cohen and Elliott (2015, Example 12.5.1), where $\Lambda^{2,0} \neq \Lambda^2$. To this end, let B denote a real-valued standard Brownian motion. Set $M = (B, -B)^\mathsf{T}$ and note that M is an $\mathbb{R}^2$-valued continuous martingale with covariation

$$[M]_t = \begin{pmatrix} t & -t \\ -t & t \end{pmatrix}, \quad \text{hence} \quad \pi \equiv \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

where we choose $\mu = \lambda$.

π is positive semi-definite, since for each $\eta \in \mathbb{R}^2$, $\eta^\mathsf{T}\pi\eta = (\eta_1 - \eta_2)^2$, which is zero precisely when $\eta_1 = \eta_2$, and positive otherwise. Let $f\colon [0,T] \to \mathbb{R}$ be measurable and such that $f \notin L^2(\lambda)$. Consider the function $g\colon [0,T] \to \mathbb{R}^2$ given by $g = (f, f)^\mathsf{T}$. By construction, we then have $g \notin \Lambda^{2,0}$, but since $\|g\|_{\Lambda^2} = 0$, we have $g \in \Lambda^2$.

For $f \in \Lambda^2$, we have $f \in L^2(M)$ in the sense of vector stochastic integration (see also Lemma 1.51 in Subsection 1.5.2). Since $[f^\mathsf{T} \bullet M]_T = \|f\|_{\Lambda^2}^2$ is deterministic and finite, Novikov's criterion shows that $Z = \mathcal{E}(f^\mathsf{T} \bullet M)$ is a strictly positive uniformly integrable martingale. Girsanov's theorem shows that under the measure $\mathbb{Q}$ with $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} = Z_t$ on $\mathcal{F}_t$ for each $t \in [0,T]$, the finite variation part in the semimartingale decomposition of M is

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

given by $[f^\mathsf{T} \bullet M, M] = h$, where $h(t) = (\mathbb{1}_{[0,t]}\pi f) \cdot \mu$ for $t \in [0,T]$. These considerations motivate the following definition.

Definition 1.13. We denote by H the set of all $h\colon [0,T] \to \mathbb{R}^d$ with the representation

$$h(t) = J(f_h)(t) \coloneqq \int_0^t \pi(s) f_h(s)\, \mu(\mathrm{d}s), \quad t \in [0,T], \tag{1.2}$$

for some $f_h \in \Lambda^2$, where the integral in (1.2) is to be understood componentwise as a Lebesgue–Stieltjes integral.

As Proposition 1.18 below will show, upon being endowed with an appropriate inner product, H becomes the Cameron–Martin space of the Gaussian measure γ_M which is induced by M on $C_0([0,T];\mathbb{R}^d)$. To the best of our knowledge, there exists no explicit characterization of the Cameron–Martin space of γ_M at the present level of generality in the literature so far, as one usually assumes M to be a Brownian motion, which is a special case of our setting (see Example 1.7).

Example 1.14. In the context of Example 1.7, H coincides with the set of absolutely continuous functions whose densities are square-integrable with respect to λ.

Remark 1.15. The Cameron–Martin space of fractional Brownian motion is not contained in our framework, except for the special case of a Brownian motion. The matrix-valued function π is not to be confused with the square-integrable but singular kernel which appears in integral representations of fractional Brownian motion and, more generally, Volterra type Gaussian processes. However, our framework can be extended to multivariate versions of these processes with representations of the form $\tilde{M}_t = \int_0^T k(t,s)\, \mathrm{d}M_s$, where k denotes an $\mathbb{R}^{d\times d}$-valued kernel function, for which, under suitable assumptions on k, the corresponding Cameron–Martin space consists of functions of the form

$$\tilde{h}(t) = \int_0^T k(t,s)\pi(s) f_h(s)\, \mu(\mathrm{d}s), \quad t \in [0,T],$$

for some $f_h \in \Lambda^2$. This formulation gives rise to the study of refined versions of multivariate Volterra type Gaussian processes as well as multivariate fractional stochastic volatility models, as one can now distinguish more explicitly between time-inhomogeneous volatility patterns which are induced by μ (or equivalently, by the quadratic variation C), the dependency structure of the components of M, which is modeled by the function π, and the path irregularities of $\tilde{M}$, which are induced by the matrix-valued kernel k.

Remark 1.16. Eq. (1.2) suggest a generalization, where the functions f_h assume values in a (possibly infinite-dimensional) Hilbert space $\tilde{H}$, and π assumes μ-almost everywhere values in the set of positive semi-definite operators on $\tilde{H}$. In this case, the integral in (1.2) is to be understood as a Lebesgue–Stieltjes–Bochner integral.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Example 1.17. Let $w\colon [0,T] \to [1,\infty)$ be a function in $L^1(\lambda;\mathbb{R})$ that is non-decreasing, set $\Sigma = I_d$, and define the measure μ through $\mu(A) = \int_A w(s)\,\lambda(\mathrm{d}s)$ for $A \in \mathcal{B}_{[0,T]}$. In line with Example 1.9, we can then construct a process M by means of the increasing and continuous function $f\colon [0,T] \to \mathbb{R}_+$ given by $f(t) \coloneqq \mu([0,t])$ for $t \in [0,T]$. In the context of Definition 1.13, the corresponding space H then has similarities to the forward curve space H_w (cf. Filipović (2001, Chapter 5)) that is used in interest rate modelling.

Since elements from H will be precisely those which we consider for the drift adjustment of M in Sections 1.3 and 1.4, we need to collect some useful properties which will be needed later on (in particular for Theorem 1.24). The following proposition collects these properties and further deepens the connections to the process M. As it turns out, being endowed with a suitable inner product, H is not only the isometric image of the space Λ^2 whose definition was inspired by the representation (1.1) of $[M]$, but H is also the Cameron–Martin space of the Gaussian measure γ_M which is induced by M on $C_0([0,T];\mathbb{R}^d)$.

Proposition 1.18. *Consider the mapping $\langle\cdot,\cdot\rangle_H$ given by $\langle g,h\rangle_H \coloneqq \langle f_g, f_h\rangle_{\Lambda^2}$ for $g,h \in H$. Then:*

(a) The integral in (1.2) is well defined for all $f_h \in \Lambda^2$ and $t \in [0,T]$;

(b) $(H, \langle\cdot,\cdot\rangle_H)$ is a real separable Hilbert space;

(c) $J\colon \Lambda^2 \to H$ is a linear isometry, and $(H^0, \langle\cdot,\cdot\rangle_H)$ is an inner product space whose completion is $(H, \langle\cdot,\cdot\rangle_H)$, where we set $H^0 \coloneqq J(\Lambda^{2,0}) \subset H$;

(d) To each $F \in H^*$ there corresponds a unique function $g_F \in H$, such that

$$F(h) = \langle g_F, h\rangle_H = \int_0^T f_{g_F}^{\mathsf{T}}(s)\pi(s)f_h(s)\,\mu(\mathrm{d}s), \quad h \in H,$$

and $\|F\|_{\mathrm{op}} = \|g_F\|_H$. Therefore, H^* is isometrically isomorphic to H;

(e) H is the Cameron–Martin space of the centered Gaussian measure γ_M that is induced by M on $E = C_0([0,T];\mathbb{R}^d)$.

Remark 1.19. Unless $\overline{H} = E$, where the closure is taken in E, the measure γ_M will be degenerate (see Remark 1.44). The identity $\overline{H} = E$ holds in some special cases, e.g. if $\mu = \lambda$ and $\pi \equiv I_d$, where the proof builds on the fact that continuous functions can be uniformly approximated by piecewise linear functions.

For the purpose of the next result, we introduce the function $I\colon E \to \overline{\mathbb{R}}_+$,

$$I(g) = \begin{cases} \frac{1}{2}\int_0^T f_g^{\mathsf{T}}(s)\pi(s)f_g(s)\,\mu(\mathrm{d}s) & \text{for } g \in H, \\ \infty & \text{otherwise.} \end{cases} \tag{1.3}$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Example 1.20. In the context of Example 1.8, π takes the form $\pi(t) = \Sigma(t)\Sigma^{\mathsf{T}}(t)$. On the other hand, in the context of Example 1.9, the measure μ could be the Lebesgue–Stieltjes measure that is induced by Cantor's ternary function and thus singular with respect to λ. The setting typically discussed in the literature, where M is a standard Brownian motion, does not encompass either of these examples.

In Subsection 1.3.1 we will discuss an importance sampling method that uses methods from the theory of large deviations. To this end, one needs to understand the asymptotic behavior of the scaled process $\sqrt{\varepsilon}M$ as $\varepsilon \searrow 0$. If M is a Brownian motion, then the corresponding result is referred to as Schilder's theorem (cf. Bogachev (1998, Corollary 4.9.3), Lifshits (2012, Theorem 8.3) and Stroock (2011, Theorem 8.4.1)). As a consequence of Proposition 1.18(e) and Proposition 1.48, we obtain the following result, whose novelty is the explicit characterization of the function I (also referred to as rate function) in (1.3) at the presented level of generality.

Proposition 1.21. *In the context of Proposition 1.18(e) we have, for each $F \in \mathcal{B}_E$,*

$$-\inf_{g\in F^\circ} I(g) \le \liminf_{\varepsilon\searrow 0} \varepsilon \log \mathbb{P}\big[\sqrt{\varepsilon}M \in F\big] \le \limsup_{\varepsilon\searrow 0} \varepsilon \log \mathbb{P}\big[\sqrt{\varepsilon}M \in F\big] \le -\inf_{g\in \bar{F}} I(g)$$

where the function $I\colon E \to \overline{\mathbb{R}}_+$ is specified in (1.3).

For notational convenience, we introduce the following convention.

General Convention 1.22. Henceforth we denote by D a dense subset of Λ^2.

Example 1.23. We have already encountered two admissible candidates for the set D: $\Lambda^{2,0}$ and $C([0,T];\mathbb{R}^d)$, see Lemma 1.10(c) and 1.10(d).

The following theorem helps us to identify dense subsets and subspaces of H, and shows how these relate to the topological support (see Remark 1.44 in Subsection 1.5.1) of the measure γ_M which we encountered in Proposition 1.18.

Theorem 1.24. *Let $H(D)$ denote the set of all $h \in H$ with $f_h \in D$. Then:*

(a) $H(D)$ is a dense subset of H which is separable when endowed with the subspace topology;

(b) If D is also a linear subspace of Λ^2, then:

(1) $(H(D), \langle\cdot,\cdot\rangle_H)$ is an inner product space, whose completion is H;

(2) There exists a countable orthonormal basis of H which consist of elements from $H(D)$;

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

(c) In the context of Proposition 1.18(e), the topological support of γ_M coincides with $\overline{H(D)}$, where the closure is taken in E. In other words,

$$\gamma_M(C_0([0,T];\mathbb{R}^d) \setminus \overline{H(D)}) = \mathbb{P}[M \in C_0([0,T];\mathbb{R}^d) \setminus \overline{H(D)}] = 0,$$

hence outside a $\mathbb{P}$-null set, paths of M can be uniformly approximated by sequences from $H(D)$.

In Section 1.2, we will discuss classes of feedforward neural networks that are also dense subsets of Λ^2, thereby satisfying Convention 1.22. Together with Theorem 1.24, this will show that we can approximate any element from H, up to the isometry J, by feedforward neural networks, which will be essential for Sections 1.3 and 1.4, where we will approximate the drift adjustment of M which minimizes the variance of the Monte Carlo estimator with feedforward neural networks.

1.2. Approximation capabilities of neural networks

In this section, we study feedforward neural networks as elements of the space Λ^2 and show how they generate, under suitable assumptions on the activation function, dense subspaces of H, thereby providing a first theoretical justification for approximating the optimal drift adjustment with feedforward neural networks when studying importance sampling in Sections 1.3 and 1.4 below.

We know from Lemma 1.10(d) that $C([0,T];\mathbb{R}^d)$ is continuously embedded into $\Lambda^{2,0}$ as a dense linear subspace. Moreover, Lemma 1.10(c) shows that $\Lambda^{2,0}$ is dense in Λ^2. Consequently, every dense linear subspace D of $C([0,T];\mathbb{R}^d)$ is densely embedded into Λ^2, thereby satisfying Convention 1.22, in which case $H(D)$ is dense in H by Theorem 1.24(a). For this reason, we first focus our attention on finding dense linear subspaces of $C([0,T];\mathbb{R}^d)$ in Proposition 1.28, before relaxing the continuity assumption in Proposition 1.29 below.

Neural networks are one particular class of functions which is of interest to us. On the one hand it satisfies the required density in $C([0,T];\mathbb{R}^d)$, and on the other hand it gives rise to efficient numerical optimization procedures which have led to fascinating results in the domain of financial and actuarial mathematics in recent years. When studying neural networks, the property of being dense in a topological space is referred to as the universal approximation property (UAP; cf. Kratsios (2021, Definition 2)). Theorems which establish the density of neural networks in a topological space are referred to as universal approximation theorems (UAT).

There are many different neural network architectures for which the universal approximation property has been shown to hold in various topological spaces (cf. Cybenko (1989); Funahashi (1989); Hornik (1991); Hornik et al. (1989); Kidger and Lyons (2020); Leshno et al. (1993); Liao et al. (2003); Mhaskar and Micchelli (1992); Park and Sandberg (1991,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

1993); Pinkus (1999); Zhang et al. (1995)). For ease of presentation, we will discuss the class of feedforward neural networks, also called multilayer perceptrons or multilayer feedforward neural networks. Note that our discussion is limited to architectures that yield universal approximation theorems in $C([0,T];\mathbb{R}^d)$ and Λ^2 (see also Subsection 1.2.1 for a UAT with respect to Hölder norms). There are many architectures for which the UAP has been established in other topological spaces, but which we will not discuss in this chapter.

Definition 1.25. Given $k \in \mathbb{N}$, $l \in \mathbb{N}$ and $\psi\colon \mathbb{R} \to \mathbb{R}$, we denote by $\mathcal{NN}^d_{k,l}(\psi)$ the set of feedforward neural networks with one neuron in the input layer, d neurons with identity activation function in the output layer, k hidden layers, and at most l hidden nodes with ψ as activation function in each hidden layer (cf. Kidger and Lyons (2020, Definition 3.1)).

Remark 1.26. Functions from $\mathcal{NN}^d_{k,l}(\psi)$ can be represented as follows. Consider affine functions $W_1\colon \mathbb{R} \to \mathbb{R}^l$, $W_{k+1}\colon \mathbb{R}^l \to \mathbb{R}^d$ and $W_2, \dots, W_k\colon \mathbb{R}^l \to \mathbb{R}^l$. For $i = 1, 2, \dots, k$, denote $F_i = \psi \circ W_i$, where the activation function ψ is applied componentwise. Then, an element from $\mathcal{NN}^d_{k,l}(\psi)$ is given by $t \mapsto W_{k+1} \circ F_k \circ \cdots \circ F_1(t)$.

If the number of nodes in the hidden layers can be arbitrarily large, we write $\mathcal{NN}^d_{k,\infty}(\psi)$. Likewise, we write $\mathcal{NN}^d_{\infty,l}(\psi)$ if the number of hidden layers can be arbitrarily large. Finally, the notation $\mathcal{NN}^d_{\infty,\infty}(\psi) = \bigcup_{k\in\mathbb{N}} \mathcal{NN}^d_{k,\infty}(\psi)$ is to be understood in an analogous way. For the purpose of Example 1.31 below, let us also introduce the following notation: If A is a finite set of functions $f\colon \mathbb{R} \to \mathbb{R}$, then $\mathcal{NN}^d_{k,l}(A)$ denotes the set of feedforward neural networks, where the hidden nodes are endowed with either of the functions from A. As a special case, we then have $\mathcal{NN}^d_{k,l}(A) = \mathcal{NN}^d_{k,l}(\psi)$ for $A = \{\psi\}$.

Functions in $\mathcal{NN}^d_{1,\infty}(\psi)$ are called shallow feedforward neural networks, while functions in $\mathcal{NN}^d_{\infty,\infty}(\psi)$ are generally referred to as deep feedforward neural networks. The set $\mathcal{NN}^d_{\infty,l}(\psi) \subset \mathcal{NN}^d_{\infty,\infty}(\psi)$ of deep narrow networks, where $l \in \mathbb{N}$ is fixed, is also of special interest (cf. Kidger and Lyons (2020)).

In the context of Definition 1.25, the function ψ is sometimes also called squashing function, sigmoid function or ridge activation function. Different terms have been chosen based on the properties of ψ, which in general differ based on which topological spaces we are studying the UAP in. For the sake of simplicity, we call ψ an activation function throughout this chapter, and impose necessary properties on ψ wherever needed. In light of Lemma 1.10(d), we are particularly interested in the UAP in $C([0,T];\mathbb{R}^d)$. At this point however, we need to discuss a technicality first.

For a given Borel measure ν on $[0,T]$ and $f, g \in C([0,T];\mathbb{R}^d)$, we write $f \sim_\nu g$ if $f = g$ outside a ν-null set. Note that $\sim_\nu$ is a binary relation on $C([0,T];\mathbb{R}^d)$ which is reflexive, symmetric and transitive, hence $\sim_\nu$ is an equivalence relation. We can thus consider the quotient space $C_\nu([0,T];\mathbb{R}^d)$ of $C([0,T];\mathbb{R}^d)$ under $\sim_\nu$, on which the ν-essential supremum $\|\cdot\|_{L^\infty([0,T],\nu)}$ is a norm, making $(C_\nu([0,T];\mathbb{R}^d), \|\cdot\|_{L^\infty([0,T],\nu)})$ a

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

normed vector space. A modification of Lemma 1.10(d) shows that $C_\nu([0,T];\mathbb{R}^d)$ is continuously embedded into $\Lambda^{2,0}$ as a dense linear subspace, provided that μ is absolutely continuous with respect to ν.

Let us collect classical versions of the universal approximation theorem which are concerned with the (almost everywhere) uniform approximation of continuous functions (cf. Hornik (1991); Kidger and Lyons (2020); Leshno et al. (1993)), as we will be referring to them in the proofs of the subsequent results.

Theorem 1.27. *Given $\psi\colon \mathbb{R} \to \mathbb{R}$, consider the assumptions:*

(1) ψ is continuous, bounded and non-constant;

(2) ψ is continuous and nonaffine, and there exists a point $x \in \mathbb{R}$ at which ψ is continuously differentiable with $\psi'(x) \neq 0$;

(3) ψ is locally λ-essentially bounded. Moreover, ψ is λ-almost everywhere not an algebraic polynomial, and the set of points of discontinuity of ψ is a λ-null set.

Then:

(a) If 1.27(1) holds, then $\mathcal{NN}^d_{1,\infty}(\psi)$ is dense in $C([0,T];\mathbb{R}^d)$;

(b) If 1.27(2) holds, then $\mathcal{NN}^d_{\infty,d+3}(\psi)$ is dense in $C([0,T];\mathbb{R}^d)$;

(c) If 1.27(3) holds, then $\mathcal{NN}^d_{1,\infty}(\psi)$ is dense in $C_\lambda([0,T];\mathbb{R}^d)$.

The following two Propositions 1.28 and 1.29 yield dense subsets of Λ^2 which consist of feedforward neural networks. We can therefore consider these sets as admissible for the set D in the context of Convention 1.22. Consequently, due to Theorem 1.24 and under suitable assumptions on ψ, feedforward neural networks are, up to the isometry J, dense in H.

Proposition 1.28 below is a consequence of Theorem 1.27 and Lemma 1.10(d). For simplicity, we only formulate it for $\mathcal{NN}^d_{1,\infty}(\psi)$, where the case for $\mathcal{NN}^d_{\infty,d+3}(\psi)$ can be argued analogously. Since $\mathcal{NN}^d_{1,\infty}(\psi)$ is a subset of $\mathcal{NN}^d_{k,\infty}(\psi)$ for every $k \in \mathbb{N}$, Propositions 1.28 and 1.29 below do hold for $\mathcal{NN}^d_{k,\infty}(\psi)$, $k \in \mathbb{N}$, too.

Proposition 1.28. *In the context of Theorem 1.27, assume either that Condition 1.27(1) holds, or that Condition 1.27(3) holds and μ is absolutely continuous with respect to λ. Then $\mathcal{NN}^d_{1,\infty}(\psi)$ is a dense linear subspace of $\Lambda^{2,0}$.*

By looking at the proof of Proposition 1.28 (which is presented in Subsection 1.5.2) it becomes clear that we cannot impose Assumption 1.27(3) in case that μ is not absolutely continuous with respect to λ. This is relevant in particular for Example 1.9, where we would need to impose either Assumption 1.27(1) or Assumption 1.27(2). Moreover, assuming ψ to be (λ-almost everywhere) continuous is also rather restrictive, given that

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

functions in Λ^2 need not be continuous. By arguing along the lines of Cybenko (1989); Hornik (1991), we can actually drop the continuity assumption on ψ, at the cost of requiring boundedness, which is not required in Assumptions 1.27(2) and 1.27(3).

Proposition 1.29. *If ψ is bounded, measurable and non-constant, then $\mathcal{NN}^d_{1,\infty}(\psi)$ is a dense linear subspace of $\Lambda^{2,0}$.*

For notational simplicity, let us convene that the activation function ψ satisfies sufficient conditions such that either Proposition 1.28 or 1.29 is applicable.

General Convention 1.30. Henceforth we assume that either Condition 1.27(1), 1.27(3) or the assumptions from Proposition 1.29 hold, depending on whether μ is absolutely continuous with respect to λ or not, and whether we need to require ψ to be (λ-almost everywhere) continuous.

Example 1.31. Fix $d = 1$ as well as $\mu = \lambda$. Note that in this case we have $\pi \equiv 1$. Set $D = \mathcal{NN}^1_{1,\infty}(\psi) = \operatorname{span}\{[0,T] \ni t \mapsto \psi(\alpha t + \eta) : \alpha, \eta \in \mathbb{R}\}$, where $\psi = \tanh$. Since every $f \in D$ is continuous and thus bounded on $[0,T]$, we may replace the Lebesgue by the Riemann integral.

If $\alpha = 0$ and $\eta \in \mathbb{R}$, then $\int_0^t \psi(\eta)\,\mathrm{d}s = \psi(\eta)t$ for $t \in [0,T]$. On the other hand, if $\alpha \neq 0$ and $\eta \in \mathbb{R}$ then, by substitution,

$$\int_0^t \psi(\alpha s + \eta)\,\mathrm{d}s = \frac{1}{\alpha}(\tilde{\psi}(\alpha t + \eta) - \tilde{\psi}(\eta)), \quad t \in [0,T],$$

where $\tilde{\psi}(\cdot) = \log(\cosh(\cdot))$. Similarly, if ψ is the standard sigmoid (logistic) function, then the same applies with $\tilde{\psi}(\cdot) = \log(1 + \exp(\cdot))$, which is also called softplus function. To sum up, we see that

$$H(D) = \operatorname{span}\{\mathrm{id}\colon [0,T] \ni t \mapsto t, \mathcal{NN}^1_{1,\infty}(\tilde{\psi})\} = \mathcal{NN}^1_{1,\infty}(\{\mathrm{id}, \tilde{\psi}\}).$$

Provided that ψ is Riemann integrable, Example 1.31 shows that, compared to the set $D = \mathcal{NN}^1_{1,\infty}(\psi)$, the set $H(D)$ can be obtained by modifying the activation function, and adding the identity function into the set of admissible activation functions. This can be helpful when optimizing over functions in $H(D)$, because one avoids having to implement integral operations. What is more, functions in $H(D)$ enjoy the property of being absolutely continuous, provided that μ is absolutely continuous with respect to λ, while this is not always the case for functions from $\mathcal{NN}^1_{1,\infty}(\psi)$, e.g. if ψ is not continuous.

Note that we formulated Propositions 1.28 and 1.29 for shallow feedforward neural networks. However, as already mentioned above, Propositions 1.28 and 1.29 do hold for the set of deep neural networks, too. For a discussion on the topic of depth vs. width, see for example Lu et al. (2017); Ronen and Ohad (2016).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.2.1. Interlude: Universal approximation in Hölder norm

In Theorem 1.27, we cited classical versions of the universal approximation theorem which are concerned with (almost everywhere) uniform approximation of continuous functions. Note that this is, in essence, a topological statement, and we may seek for refined approximation results that hold with respect to stricter topologies. Natural candidate topologies with respect to which we may seek to derive a universal approximation theorem are Hölder type topologies.

Given $\alpha \in (0,1)$, we denote by $E_\alpha = C_0^\alpha([0,T];\mathbb{R}^d) \subset C_0([0,T];\mathbb{R}^d)$ the vector space of $\mathbb{R}^d$-valued, α-Hölder continuous functions on $[0,T]$ that are zero at the origin. The space E_α is also referred to as α-Hölder space. We endow this space with the topology which is induced by the norm

$$E_\alpha \ni f \mapsto \|f\|_\alpha := \sup_{\substack{s,t\in[0,T]\\ 0<t-s\leq 1}} \frac{|f(t)-f(s)|}{(t-s)^\alpha}.$$

The Kolmogorov–Chentsov continuity theorem shows that all paths of $\mathbb{R}^d$-valued standard Brownian motion are α-Hölder continuous for every $\alpha \in (0,1/2)$, hence it would be desirable to use E_α as the space on which to consider the restriction of the classical Wiener measure (see Example 1.1 and Definition 1.42). Although $(E_\alpha, \|\cdot\|_\alpha)$ is indeed a real Banach space, it does not contain a countable dense subset and is thus not separable (cf. Schmock (2024, Exercise 2.126)).

We need to pass to the little α-Hölder space, i.e. the subspace $E_{\alpha,0}$ of all $f \in E_\alpha$ that satisfy $|f(t)-f(s)| = o(|t-s|^\alpha)$ as $|t-s| \searrow 0$. Then $(E_{\alpha,0}, \|\cdot\|_\alpha)$ is a real Banach space. $E_{\alpha,0}$ is also referred to as the space of α-Hölder paths with vanishing Hölder oscillation (cf. Friz and Hairer (2020, Exercise 2.12)). Note that $E_{\alpha,0}$ has a very useful characterization: It is the closure of $C_0^\infty([0,T];\mathbb{R}^d)$, the vector space of $\mathbb{R}^d$-valued smooth functions on $[0,T]$ that are zero at the origin, where the closure is taken with respect to the topology induced by $\|\cdot\|_\alpha$. Moreover, we have the inclusion $E_\beta \subset E_{\alpha,0}$ for all $0 < \alpha < \beta < 1$.

In the context of Gaussian measures and large deviations theory, the space $E_{\alpha,0}$ has been studied in great detail (cf. Andresen et al. (2013); Baldi et al. (1992); Ciesielski (1960)). In particular, the following important property has been shown to hold: If we fix $\pi = I_d$ and $\mu = \lambda$, then H is continuously embedded into $E_{\alpha,0}$ for each $\alpha \in (0,1/2)$, and there exists a countable family of functions in H (the Faber–Schauder system) that constitutes a Schauder basis of $(E_{\alpha,0}, \|\cdot\|_\alpha)$. While on the one hand this implies the separability of $E_{\alpha,0}$, more importantly, we see that H is not only continuously, but also densely embedded into $E_{\alpha,0}$. In conjunction with Theorem 1.24, and as a direct consequence to this observation, we obtain the following result.

Proposition 1.32. *Fix $\mu = \lambda$, $\pi = I_d$ and $\alpha, \beta \in (0,1/2)$ with $\beta < \alpha$. Then, in the*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

context of Propositions 1.28 and 1.29, for each $f \in E_{\alpha,0}$, there exists a sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{NN}^d_{1,\infty}(\psi)$ such that

$$\lim_{n\to\infty} \|f - J(f_n)\|_\alpha = 0.$$

In particular, every smooth function $f \in C_0^\infty([0,T];\mathbb{R}^d)$ and every α-Hölder continuous function $f \in E_\alpha$ can be approximated, up to the linear isometry J (see Definition 1.13), by sequences from $\mathcal{NN}^d_{1,\infty}(\psi)$ with respect to $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, respectively.

Let us conclude this subsection with several remarks. First, note that, based on Ciesielski et al. (1993), Proposition 1.32 should extend to certain Besov–Orlicz type norms, which induce stricter topologies than the Hölder norms. Moreover, it should be possible to relax the assumption $\mu = \lambda$ by considering a modified Hölder norm with denominator $\mu((s,t])^\alpha$ instead of $(t-s)^\alpha$. Finally, note that the universal approximation property of neural networks in topological spaces with topologies that are stricter than the one induced by the uniform norm have already been studied in the literature. See e.g. Gühring et al. (2020), which studies the UAP in Sobolev spaces.

1.3. Importance sampling with feedforward neural networks

Upon having studied the tractable space H of drift adjustments which coincides with the Cameron–Martin space of the Gaussian measure γ_M, and having proved that feedforward neural networks are, up to the isometry J, dense in H by combining Theorem 1.24 and Propositions 1.28 resp. 1.29, we now turn our attention to importance sampling. To this end, we first write down the basic setting. In Subsection 1.3.1, we then study how our method complements a classical approach which employs ideas from large deviations theory, before finally studying the full problem in Subsection 1.3.2. Most notably, Theorem 1.38 below provides a theoretical justification for our simulations in Section 1.4, see also Remark 1.39.

Let $\mathcal{C}$ denote the vector space of $\mathbb{R}^n$-valued, continuous and $\mathbb{F}$-adapted processes. Let $a\colon \Omega \times [0,T] \times \mathcal{C} \to \mathbb{R}^n$ and $b\colon \Omega \times [0,T] \times \mathcal{C} \to \mathbb{R}^{n\times d}$ be non-anticipative coefficients (cf. Cohen and Elliott (2015, Definition 16.0.3)), such that the stochastic differential equation

$$\mathrm{d}X_t = a_t(X)\,\mathrm{d}C_t + b_t(X)\,\mathrm{d}M_t, \quad t \in [0,T], \tag{1.4}$$

subject to $X_0 \equiv x \in \mathbb{R}^n$, admits a unique weak solution. For ease of notation, we write $a_t(X)$, $b_t(X)$ instead of $a(\omega,t,X)$, $b(\omega,t,X)$, and understand Eq. (1.4) to hold componentwise, i.e. $X^i_t = x^i + (a(X)_i \bullet C)_t + (b(X)_{i,\cdot} \bullet M)_t$ for each $i \in \{1,2,\dots,n\}$ and $t \in [0,T]$.

Let F be a real-valued random functional on $\Omega \times C([0,T];\mathbb{R}^n)$, such that the mapping $\Omega \ni \omega \mapsto F(\omega, X(\omega))$ is $\mathcal{F}_T$-measurable. For simplicity we write $F(\cdot, X) = F(X)$, and

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

call $F(X)$ a random payoff. We are interested in obtaining a Monte Carlo estimate of its expectation under $\mathbb{P}$,

$$\mathbb{E}_{\mathbb{P}}[F(X)] = \int_{\Omega} F(\omega, X(\omega))\mathbb{P}(\mathrm{d}\omega), \tag{1.5}$$

provided that (1.5) is a real number.

Remark 1.33. If $\mathbb{E}_{\mathbb{P}}[F(X)]$ is to denote an option price, then we would require $\mathbb{P}$ to be a risk-neutral measure. However, the results in this section do not require $\mathbb{P}$ to be risk-neutral. Actually, we do not need to assume $\mathbb{E}_{\mathbb{P}}[F(X)]$ to be an option price, as long as X follows the SDE (1.4) and $F(X) \in \mathcal{L}^0(\mathcal{F}_T)$.

Example 1.34. The SDE (1.4) can model the evolution of asset prices within stochastic volatility models. Thus, our method complements Robertson (2010), where the author employs methods from the theory of large deviations in order to derive asymptotically optimal drift adjustments (see Subsection 1.3.1, where we discuss asymptotic optimality in the context of Guasoni and Robertson (2008)) for pricing stochastic volatility models, very much in the spirit of Guasoni and Robertson (2008).

Let $(X^i)_{i\in\mathbb{N}}$ denote a sequence of independent copies of solutions to (1.4). By the strong law of large numbers, the sample means $Z_k = \sum_{i=1}^k F(X^i)/k$ converge $\mathbb{P}$-almost surely to $m = \mathbb{E}_{\mathbb{P}}[F(X)]$. Moreover, if $F(X^1)$ has a finite variance $\sigma^2 > 0$ then, according to the central limit theorem, as $k \to \infty$, the law of $\sqrt{k}(Z_k - m)$ converges weakly to $\mathcal{N}(0, \sigma^2)$. We therefore see that $Z_k - m$ is approximately normally distributed with mean zero and standard deviation $\sigma/\sqrt{k}$. In practice, the standard error $\sigma/\sqrt{k}$ can be quite large even for large values of k, which calls for the application of variance reduction methods.

Remark 1.35. For each $f \in \Lambda^2$, we have that $f^{\mathsf{T}} \bullet M$ is a real-valued continuous local martingale with $[f^{\mathsf{T}} \bullet M]_t = \int_0^t f^{\mathsf{T}}(s)\pi(s)f(s)\,\mu(\mathrm{d}s)$ for $t \in [0, T]$. As a consequence, similarly as argued in Remark 1.4, $f^{\mathsf{T}} \bullet M$ is a Gaussian $\mathbb{F}$-Markov process with

$$\mathcal{L}\Big((f^{\mathsf{T}} \bullet M)_t - (f^{\mathsf{T}} \bullet M)_s\Big) = \mathcal{N}\Big(0, \int_s^t f^{\mathsf{T}}(u)\pi(s)f(u)\mu(\mathrm{d}u)\Big),$$

for $s < t$ in $[0, T]$, which shows how one can simulate increments of $f^{\mathsf{T}} \bullet M$, provided that the integrals $\int_s^t f^{\mathsf{T}}(u)\pi(s)f(u)\mu(\mathrm{d}u)$ can be explicitly computed.

Recall that, whenever Y is a real-valued continuous semimartingale, the Doléans exponential $\mathcal{E}(Y)$ is the strictly positive continuous semimartingale that is, up to indistinguishability, the unique solution in $L(Y)$ to the stochastic integral equation

$$\mathcal{E}(Y) = \exp(Y_0) + \int_0^{\cdot} \mathcal{E}(Y)_s \,\mathrm{d}Y_s,$$

and is given by $\mathcal{E}(Y) = \exp(Y - [Y]/2)$ (cf. Schmock (2024, Theorem 6.61)). By a

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

generalization of the functional equation of the exponential function, $\mathcal{E}(Y)^{-1} = \mathcal{E}(-Y + [Y])$. We refer to Rheinländer (2010) for a survey on stochastic exponentials.

Since M is a continuous local martingale and $[f_h^\mathsf{T} \bullet M]_T = \|h\|_H^2 < \infty$ for each $h \in H$, $f_h^\mathsf{T} \bullet M$ is a square integrable martingale (see also Lemma 1.51 in Subsection 1.5.2), and $\mathcal{E}(f_h^\mathsf{T} \bullet M)$ is a non-negative continuous local martingale, hence a supermartingale. Moreover, since $\mathbb{E}_\mathbb{P}[\exp(\frac{k}{2}[f_h^\mathsf{T} \bullet M]_T)] = \exp(\frac{k}{2}\|h\|_H^2) < \infty$ for every $k > 1$, Cohen and Elliott (2015, Theorem 15.4.6) and Lépingle and Mémin (1978) show that $\mathcal{E}(f_h^\mathsf{T} \bullet M) \in \mathcal{H}^p$ for every $p = k/(2\sqrt{k} - 1) > 1$ with upper bound

$$\|\mathcal{E}(f_h^\mathsf{T} \bullet M)\|_{\mathcal{H}^p} \leq \frac{p}{p-1} \exp\Big(\frac{\sqrt{k}-1}{2}\|h\|_H^2\Big).$$

An application of de la Vallée Poussin's criterion (see e.g. Cohen and Elliott (2015, Corollary 2.5.5)) further shows that $\mathcal{E}(f_h^\mathsf{T} \bullet M) \in \mathcal{H}^p$ implies the uniform integrability of $\mathcal{E}(f_h^\mathsf{T} \bullet M)$.

By a change of measure, Eq. (1.5) can now be rewritten as

$$\mathbb{E}_\mathbb{P}[F(X)] = \mathbb{E}_{\mathbb{P}_h}[F(X)(\mathcal{E}(f_h^\mathsf{T} \bullet M)^{-1})_T], \tag{1.6}$$

where $\mathbb{P}_h$ is defined by $\mathrm{d}\mathbb{P}_h = \mathcal{E}(f_h^\mathsf{T} \bullet M)_t \,\mathrm{d}\mathbb{P}$ on $\mathcal{F}_t$ for all $t \in [0, T]$. If we denote by $F_h(X) \coloneqq F(X)(\mathcal{E}(f_h^\mathsf{T} \bullet M)^{-1})_T$ the modified random payoff, then the $\mathbb{P}$-expectation of $F(X)$ and the $\mathbb{P}_h$-expectation of $F_h(X)$ are identical. Provided that $F(X)$ has a finite second moment with respect to $\mathbb{P}$, the variance of $F_h(X)$ under $\mathbb{P}_h$ is given by

$$\mathbb{E}_{\mathbb{P}_h}[F_h^2(X)] - \mathbb{E}_{\mathbb{P}_h}[F_h(X)]^2 = \mathbb{E}_\mathbb{P}[F^2(X)(\mathcal{E}(f_h^\mathsf{T} \bullet M)^{-1})_T] - \mathbb{E}_\mathbb{P}[F(X)]^2. \tag{1.7}$$

Therefore, we can compute (1.6) under the measure $\mathbb{P}_h$ and try to find $h \in H$ such that (1.7) is minimized. Note that the second term on the right-hand side of (1.7) does not depend on h, so that we focus on minimizing the first term, which for each $h \in H$ is given by

$$V(h) \coloneqq \mathbb{E}_\mathbb{P}[F^2(X)(\mathcal{E}(f_h^\mathsf{T} \bullet M)^{-1})_T] = \mathbb{E}_\mathbb{P}\big[F^2(X)\exp\big(-(f_h^\mathsf{T} \bullet M)_T + \|h\|_H^2/2\big)\big].$$

To sum up, our problem reads

$$\min_{h \in H} V(h), \tag{1.8}$$

provided that a minimizer of V exists (see Theorem 1.38 for sufficient conditions).

1.3.1. Approximating the asymptotically optimal sampling measure

Before we turn our attention to solving (1.8), let us first discuss the classical approach presented in Guasoni and Robertson (2008), that uses methods from the theory of large deviations, and show how our method complements it. Note that the setting of Guasoni

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

and Robertson (2008) is a special case of the setting of Section 1.3. To this end, set $d = 1$, let M be a standard Brownian motion and let $\mathbb{F}$ be the augmented natural filtration of M (which satisfies the usual hypotheses). Set $X = M$ and assume that the payoff $F\colon C_0([0,T];\mathbb{R}) \to \mathbb{R}_+$ is continuous, where $C_0([0,T];\mathbb{R})$ is endowed with the topology of uniform convergence.

In Guasoni and Robertson (2008), the authors argue that (1.8) is in general intractable. Rather than minimizing (1.8), the authors consider for each $h \in H$ the small-noise limit

$$L(h) := \limsup_{\epsilon \searrow 0} \epsilon \log \mathbb{E}_{\mathbb{P}}\Big[\exp\Big(\frac{1}{\epsilon}\big(2\tilde{F}(\sqrt{\epsilon}M) - ((\sqrt{\epsilon}f_h) \bullet M)_T + \|h\|_H^2/2\big)\Big)\Big], \tag{1.9}$$

where $\tilde{F} := \log F$. The limit (1.9) corresponds to approximating $V(h) \approx \exp(L(h))$.

Assume that $\tilde{F}\colon C_0([0,T];\mathbb{R}) \to \mathbb{R} \cup \{-\infty\}$ is continuous. Moreover, assume that there exist constants $K_1, K_2 > 0$ as well as $\alpha \in (0,2)$ such that $\tilde{F}(x) \leq K_1 + K_2\|x\|_\infty^\alpha$ for each $x \in C_0([0,T];\mathbb{R})$. As Guasoni and Robertson (2008, Theorem 3.6) shows, one can invoke a version of Varadhan's integral lemma to rewrite (1.9) as a variational problem, provided that h is an element of H_{bv}, the space of all $h \in H$ such that $f_h \in \Lambda^2 = L^2(\lambda)$ is of bounded variation, and aim to solve $\min_{h \in H_{\mathrm{bv}}} L(h)$, provided that a minimizer exists.

For the proof of the central result Guasoni and Robertson (2008, Theorem 3.6), the following functional is important: For $M > 0$ and $h \in H$ let $\tilde{F}_{h,M}\colon H \ni g \mapsto 2\tilde{F}(g) - M\|g+h\|_H^2 + \|h\|_H^2$. By Guasoni and Robertson (2008, Lemma 7.1), there exists a maximizer $g_{h,M} \in H$ of $\tilde{F}_{h,M}$. Together with Proposition 1.28, we then obtain

Proposition 1.36. *Assume that $\psi\colon \mathbb{R} \to \mathbb{R}$ is continuously differentiable, bounded and non-constant, and set $D = \mathcal{NN}^1_{1,\infty}(\psi)$. Then $H(D) \subset H_{\mathrm{bv}}$, and for each $M > 0$ and $h \in H$, there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $H(D)$ such that*

$$\lim_{n\to\infty} \tilde{F}_{h,M}(h_n) = \tilde{F}_{h,M}(g_{h,M}) = \max_{g \in H}\big(2\tilde{F}(g) - M\|g+h\|_H^2 + \|h\|_H^2\big).$$

According to Guasoni and Robertson (2008, Theorem 3.6) and the discussion thereafter, the strategy for finding a minimizer of L is as follows: Find a maximizer $g_{h,1}$ to $\tilde{F}_{h,1}$ for $h \equiv 0$. Verify whether $g_{h,1}$ is actually an element of H_{bv} and, if this is the case, then $g_{h,1}$ minimizes L provided that $L(g_{h,1}) = \tilde{F}_{h,1}(g_{h,1})$ holds true. Since the evaluation of L at $g_{h,1}$ involves having to find a maximizer of $\tilde{F}_{g_{h,1},1/2}$, checking whether $L(g_{h,1}) = \tilde{F}_{h,1}(g_{h,1})$ might only be feasible by a numerical approximation which introduces an error. However, if one could establish said identity, then $g_{h,1}$ would be a minimizer of L, in which case we say that $g_{h,1}$ is asymptotically optimal.

Let us assume that there exists a maximizer $g_{h,1}$ to $\tilde{F}_{h,1}$ for $h \equiv 0$ that is indeed asymptotically optimal. In light of Proposition 1.36, we can then approximate $g_{h,1}$ by a sequence $(h_n)_{n\in\mathbb{N}}$ from $H(\mathcal{NN}^1_{1,\infty}(\psi))$ such that $\tilde{F}_{h,1}(h_n)$ converges to $\tilde{F}_{h,1}(g_{h,1})$. Theorem 1.38(b) below then implies that $V(h_n)$ converges to $V(g_{h,1})$. To sum up, rather

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

then trying to find a minimizer of V, one might instead study

$$\max_{g\in H} \tilde{F}_{h,1}(g) = \max_{g\in H} \left(2\tilde{F}(g) - \|g\|_H^2\right) \tag{1.10}$$

and solve the modified problem (1.10) with feedforward neural networks.

1.3.2. Approximating the optimal sampling measure

In what follows, we consider the full problem (1.8) and propose to solve it with feedforward neural networks. One advantage of this approach is that we do not, in contrast to Subsection 1.3.1, seek to find an asymptotically optimal drift adjustment by minimizing (1.9), but rather stay within the full problem (1.8). Moreover, Theorem 1.38(d) and Remark 1.39 provide a theoretical justification for employing the tractable class of shallow feedforward neural networks for this optimization problem. The numerical simulations in Section 1.4 will demonstrate that indeed, we obtain substantial reductions in the variance of the Monte Carlo estimators for several multivariate asset price processes and path-dependent payoff functionals.

In the following lemma, we consider a nonlinear operator which maps elements from Cameron–Martin space to probability densities. This result is essential, as it will imply in Theorem 1.38 below that the optimal sampling measure can be approximated by measures which are generated by feedforward neural networks.

Lemma 1.37. *The operator* $A_p\colon H \ni h \mapsto (\mathcal{E}(f_h^\mathsf{T} \bullet M)^{-1})_T \in L^p(\mathbb{P})$ *is continuous for each* $p \in [1,\infty)$. *Moreover,* A_p *is not quasi-bounded, meaning that*

$$\limsup_{\|h\|_H\to\infty} \frac{\|A_p(h)\|_{L^p(\mathbb{P})}}{\|h\|_H} = \infty.$$

Finally, we formulate Theorem 1.38. The proof (which is presented in Subsection 1.5.2) complements Lemaire and Pagès (2010, Proposition 4) and not only shows, under rather weak assumptions, that the functional V does indeed admit a minimizer, but Theorem 1.38(d) applied to a dense subset D of Λ^2 which consists of feedforward neural networks provides the theoretical justification for the simulations in Section 1.4, see also Remark 1.39 below.

Theorem 1.38. *Assume that* $\mathbb{P}[F^2(X) > 0] > 0$, *and that there exists some* $\varepsilon > 0$ *such that* $F(X) \in L^{2+\epsilon}(\mathbb{P})$. *Then:*

(a) V *is* $\mathbb{R}_+$*-valued;*

(b) V *is continuous;*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

(c) There exists a minimizer of V, i.e.

$$\arg\min_{h\in H} V(h) = \{g \in H : V(g) \leq V(h), \forall h \in H\} \neq \varnothing;$$

(d) There exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $H(D)$ such that $\lim_{n\to\infty} V(h_n) = \min_{h\in H} V(h)$.

Remark 1.39. In the context of Theorem 1.38(d), we may seek to find a minimizer of V by performing measure changes which are induced by Doléans exponentials of the form $\mathcal{E}(f^{\mathsf{T}} \bullet M)$, where $f \in \mathcal{NN}^d_{1,\infty}(\psi)$. In Section 1.4, we pursue this approach for several different asset price models, achieving substantial reductions in the variance of the corresponding Monte Carlo estimators.

Remark 1.40. Theorem 1.38(d) shows that neural network-induced changes of the sampling measure can approximate the optimal sampling measure arbitrarily well in the sense that the second moment of the modified payoff under the optimal measure can be approximated up to an arbitrarily small $\epsilon > 0$. However, the proof is not constructive, it does not deliver a recipe how to actually obtain such a sequence (h_n) of neural network-induced elements from Cameron–Martin space that converges to the optimum. In Section 5 below, we use stochastic gradient descent to train our neural networks. This procedure builds on the method of stochastic approximation, which was pioneered in Robbins and Monro (1951). Stochastic approximation for importance sampling for option pricing in continuous-time models has been studied in Lemaire and Pagès (2010). We refer to their Section 3 for details on how to construct convergent sequences of functions based on the method of stochastic approximation.

Remark 1.41. Let us assume that the SDE (1.4) depends on a set of parameters $\alpha \in \mathbb{R}^m$ for some $m \in \mathbb{N}$. Fix $i \in \{1, 2, \ldots, m\}$, and let us further assume that we can exchange the order of differentiation and integration, i.e. $\frac{\partial}{\partial\alpha_i}\mathbb{E}_{\mathbb{P}}[F(X)] = \mathbb{E}_{\mathbb{P}}[\frac{\partial}{\partial\alpha_i}F(X)]$. If we wanted to jointly reduce the standard error of the Monte Carlo estimators of the expected random payoff and of its sensitivity with respect to α_i, we could modify the definition of V:

$$\tilde{V}(h) = \mathbb{E}_{\mathbb{P}}\big[(w_1 F^2(X) + w_2(\frac{\partial}{\partial\alpha_i}F(X))^2)(\mathcal{E}(f_h^{\mathsf{T}} \bullet M)^{-1})_T\big], \quad h \in H,$$

where $w_1, w_2 \in (0, 1)$ are weights that sum up to 1. If there exists some $\varepsilon > 0$ such that $w_1 F^2(X) + w_2(\frac{\partial}{\partial\alpha_i}F(X))^2 \in L^{1+\varepsilon}(\mathbb{P})$ and $\mathbb{P}[w_1 F^2(X) + w_2(\frac{\partial}{\partial\alpha_i}F(X))^2 > 0] > 0$, then Theorem 1.38 applies correspondingly. Analogous considerations hold for higher-order sensitivities as well as for the joint reduction of standard errors for more than one sensitivity. We refer the reader to Glasserman (2003, Section 7.2) for details on the computation of pathwise derivatives for some classical models and payoffs.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.4. Numerical study

In this section, we provide a range of carefully chosen numerical examples to showcase the various strengths of our method. Additionally, we will compare our approach to other methods that have been proposed in the literature. All computational tasks were performed using Python, leveraging the Keras deep learning API for the construction and training of our neural networks. All codes that were used for the simulations are available on Github, see https://github.com/aarandjel/importance-sampling-with-feedforward-neural-networks.

Let us provide a brief overview of the examples appearing in the subsequent subsections. In Subsection 1.4.2, we explore a time-change instance that deviates from the conventional assumption of $\mu = \lambda$ to better represent phases of changing business activity. Subsection 1.4.3 considers a knock-out option and discusses the occurrence of multiple rare events. Moving on to Subsection 1.4.4, we examine a stochastic volatility model with an imposed dynamic correlation structure, which directly influences the norm on Cameron–Martin space. Lastly, in Subsection 1.4.5, we investigate the feasibility of utilizing neural networks for importance sampling in a high-dimensional model. Throughout all of our examples, we consider arithmetic Asian (basket) call options with strike K and basket weights w as the chosen payoffs,

$$F(X) = \Big(\frac{1}{T}\int_0^T \langle w, X_t\rangle \,\mathrm{d}t - K\Big)^+,$$

while Subsection 1.4.3 additionally incorporates knock-out barriers for further analysis.

To establish a solid basis for comparison, we have selected the methodologies proposed by Glasserman et al. (1999), Guasoni and Robertson (2008), Capriotti (2008), Arouna (2003), Su and Fu (2000), as well as Lemaire and Pagès (2010). To underscore the versatility of our approach in handling more general models than those presented in the literature, we will initially present results for the models discussed in the previous paragraph. Subsequently, we will report results from simulations performed for the models studied in the literature mentioned above.

To train a feedforward neural network, our approach is as follows. First, we simulate N trajectories X^i, $i = 1, \dots, N$ of the asset price using the Euler–Maruyama method, based on a pre-defined time-grid. Then, we decide on a set $\mathcal{NN}^d_{k,l}(\psi)$ from which we seek to identify the optimal function, by selecting the number of hidden layers k, the number of hidden nodes l, and the activation function ψ. The output dimension d of the neural networks aligns with the dimension of the process M. We approximate V by computing an average over the N trajectories,

$$V(\theta) = \frac{1}{N}\sum_{i=1}^{N} F^2(X^i)\exp(-(f_\theta^{\mathsf{T}} \bullet M^i)_T + \|f_\theta\|^2_{\Lambda^2}/2), \tag{1.11}$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

where θ represents the vector encompassing all trainable parameters of the neural network f_θ, and all quantities on the right-hand side of Eq. (1.11) are appropriately discretized. We therefore consider V as a function of the finite parameter vector θ, and aim to find the optimal θ^* and thus the optimal element f_{θ^*} from $\mathcal{NN}^d_{k,l}(\psi)$.

To achieve this, we employ stochastic gradient descent, a technique originally pioneered in Robbins and Monro (1951). Specifically, we adopt the mini-batch variant of this method, which replaces the mean over all N trajectories with means over smaller sub-batches. Starting from an initial guess, the parameter-vector θ is then iteratively updated with a scaled version of the gradient of V over those sub-batches, i.e. $\theta_{m+1} = \theta_m - \gamma_m \nabla_{batch} V(\theta_m)$ with learning rate γ_m and $\nabla_{batch} V$ denoting the gradient of V over one specific batch. Upon completing a full iteration through all batches, we consider the neural network to have completed one epoch of training. For each subsequent epoch, the trajectories contained in the individual batches can then be randomly shuffled around, and the parameter θ is updated until a stopping criterion is reached. One notable advantage of neural networks lies in their ability to efficiently compute gradients through the back-propagation method. Additionally, we utilize a popular modified version of this training routine known as Adam (cf. Kingma and Ba (2017)), which incorporates the first and second moments of the gradient estimates to enhance performance.

In all of our subsequent examples, we train the neural networks using 100 batches, each consisting of 1,024 trajectories. For validation purposes, we employ an additional 100 batches, also comprising 1,024 trajectories, and stop the training process when the loss, $V(\theta)$, ceases to reduce on the validation set. The results presented in the following tables are derived from simulations performed on separate test datasets, each containing 10^5 trajectories. Throughout the training, validation, and testing phases, we maintain a fixed learning rate of 10^{-3} for the stochastic gradient descent, and we fix the time horizon to $T = 1$ to consider the time interval $[0, 1]$. Unless otherwise specified, we utilize a step size of $\Delta t = 1/250$. However, in Subsection 1.4.5, we deviate from this convention. We employ a step size corresponding to $\Delta t = d/10^4$ during the training and validation process, where d denotes the dimension of the asset price process. For example, when $d = 200$ then $\Delta t = 1/50$. This adjustment is only implemented for dimensions ranging from $d = 100$ to $d = 1{,}000$, while the step size always remains $\Delta t = 1/100$ for the testing dataset, as well as for the training and validation datasets in case $d < 100$. In all simulations described below, we train shallow feedforward neural networks with a single hidden layer, using $\psi(x) = \tanh(x)$ as activation function. The number of hidden nodes used for the various examples is reported beneath the tables. The tables below present results for different choices of model parameters, presenting mean estimates, standard errors, relative standard errors as a percentage of the mean, and variance ratios. The variance ratios were obtained by comparing the variance of the mean estimate from both a Monte Carlo and a Monte Carlo with importance sampling run, dividing the former by the estimate of the latter.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.4.1. Stratified sampling with feedforward neural networks

In addition to importance sampling, stratified sampling is a widely used variance reduction method. Stratified sampling involves constraining the fraction of trajectories sampled from specific subsets of the sample space. To implement this method effectively, suitable subsets of the sample space need to be chosen, covering the entire sample space, along with the desired fractions of the overall sample falling within each subset. It is important to note that stratification typically generates dependent sequences of random variables, which affects the calculation of the standard error and variance of the Monte Carlo estimator. For further information on this approach, we refer to Glasserman (2003).

In Glasserman et al. (1999), the authors investigate importance sampling and stratification techniques for pricing path-dependent options. Similar to Guasoni and Robertson (2008), they employ large deviations techniques to determine asymptotically optimal drift adjustments in a discrete-time framework. In order to overcome the computational effort that might be required to perform optimal stratification, the authors propose utilizing the drift identified for importance sampling to perform further stratification. In the following examples, we will augment our results based on importance sampling with the stratified sampling approach.

More precisely, let us consider the estimation of $\mathbb{E}[F(X)]$. Having discretized the time interval into m points, assume that $F(X)$ can be expressed as a function of Z, with Z being a m-dimensional vector of independent standard normal variables. If f denotes the optimal element from Cameron–Martin space, sampled at t_i as a vector, and appropriately re-scaled such that adjusting the drift of M corresponds to adding f to Z in the discrete time case, then we want to sample Z conditional on $f^\top Z \in A_i$, where A_i denotes a stratum, which is a subset of the sample space. In our case, A_i is chosen to correspond to the interval between the $(i-1)/N$ and the i/N quantile of the standard normal distribution, where N denotes the number of strata. We maintain an equal number of replications for each stratum. For further details on simulating Z conditional on $f^\top Z \in A_i$, we refer readers to Section 4 in Glasserman et al. (1999).

In Subsections 1.4.2-1.4.4, we extend our analysis beyond importance sampling by additionally using the trained neural networks to implement stratified sampling. By combining these two techniques, we demonstrate the significant potential for further variance reduction. It is crucial to emphasize that using the optimal importance sampling drift for stratification may not always result in optimal stratification in general. Furthermore, it is worth noting that the setting of Glasserman et al. (1999) is discrete in time. There is ample scope to explore optimal stratified sampling in continuous time using neural networks.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.4.2. Changing business activity

Methods typically employed for importance sampling based on continuous stochastic processes for asset prices often assume that the dynamics of the asset price are governed by an SDE driven by a Brownian motion. Here, we aim to deviate from the conventional framework where $\mu = \lambda$, and explore an example involving a time-changed Brownian motion. It is important to note that in this case, the time-change directly affects the definition of the Cameron–Martin norm through the Lebesgue–Stieltjes measure μ. The utilization of a deterministic time-change can be interpreted as a means of modelling periods characterized by varying business activity, thus incorporating effects such as seasonality. See Li et al. (2016) for an example where this has been done.

Let us consider the asset X governed by the dynamics $\mathrm{d}X_t = rX_t\,\mathrm{d}[M]_t + \sigma X_t\,\mathrm{d}M_t$, where $X_0 = x$ and $M = B_{C_t}$ for B representing a standard Brownian motion. Motivated by Li et al. (2016), we make the assumption that $C_t = \int_0^t \nu(s)\,\mathrm{d}s$, where the activity rate function ν takes the form

$$\nu(s) = \begin{cases} 1 + \kappa(s - 0.2)/0.1, & s \in [0.2, 0.3), \\ 1 + \kappa(0.4 - s)/0.1, & s \in [0.3, 0.4), \\ 1 + 2\kappa(s - 0.6)/0.1, & s \in [0.6, 0.7), \\ 1 + 2\kappa(0.8 - s)/0.1, & s \in [0.7, 0.8), \\ 1, & \text{else}, \end{cases}$$

where κ denotes the level of business activity. Moreover, we normalize ν such that $C_1 = \int_0^1 \nu(s)\,\mathrm{d}s = 1$. In this case, μ is absolutely continuous with respect to λ with Radon–Nikodým density ν, and $[M] = C$.

Figure 1.1 illustrates a representative trajectory of X under the assumption of an activity rate function modeled by $\kappa = 10$. The trajectory exhibits two distinct phases characterized by heightened volatility, which can be interpreted as periods of increased business activity. Table 1.1 below presents results obtained for various values of κ. Note that the special case of $\kappa = 0$ corresponds to the classical Black–Scholes model.

From Table 1.1 it is evident that both importance sampling and the combined approach of importance- and stratified sampling exhibit substantial variance reduction across all values of κ. Notably, the combination of importance- and stratified sampling demonstrates a remarkable enhancement in variance reduction compared to using importance sampling alone.

In Guasoni and Robertson (2008), the authors study asymptotically optimal importance sampling in continuous time following a large deviations approach. In Table 2 of their work, the authors present variance ratios for an arithmetic Asian call option within a Black–Scholes model across various values of volatility (sigma) and strike (K). We refer to Guasoni and Robertson (2008, Section 5) for details about the model and the selected parameters. We replicated their Table 2 using neural networks to induce optimal measure

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

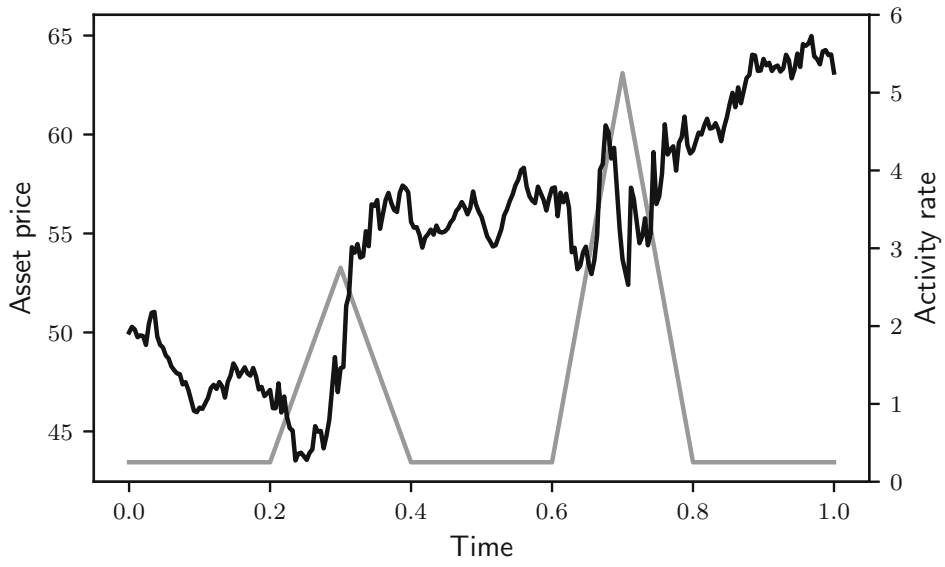

Figure 1.1. *Typical sample path for the model described above with* $\kappa = 10$*, along with the corresponding activity rate function* ν*. Other model parameters are* $X_0 = 50$*,* $r = 0.05$ *and* $\sigma = 0.25$*.*

Table 1.1. *Variance ratios for different levels* κ *of business activity.*

Param.	**Importance Sampling**			**IS and Stratification**		
κ	Mean	Std. err.	Var. ratio	Mean	Std. err.	Var. ratio
0	5.945	0.019 (0.32%)	129	5.9337	0.0018 (0.03%)	14,525
1	4.675	0.015 (0.32%)	154	4.6672	0.0023 (0.05%)	6,714
2	3.987	0.013 (0.33%)	167	3.9860	0.0026 (0.07%)	4,191
5	3.053	0.010 (0.33%)	207	3.0495	0.0016 (0.05%)	8,139
10	2.5286	0.0086 (0.34%)	235	2.5304	0.0016 (0.06%)	7,110

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes is 2. Other model parameters are $X_0 = 50$, $r = 0.05$, $\sigma = 0.25$ and $K = 70$.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

changes and subsequently compared the obtained variance ratios. On average, employing neural networks resulted in a 20% increase in the variance ratio. For instance, when considering a volatility of 30% and a strike of 70, Guasoni & Robertson report a variance ratio of 56, while our method yielded a variance ratio of 67.

In Capriotti (2008), the author studies importance sampling based on a least-squares optimization procedure. The author presents variance ratios for various combinations of volatility σ and strike K in Table 6, specifically for an arithmetic Asian call option within a Black–Scholes model. Additionally, the table includes variance ratios obtained using an adaptive Robbins–Monro procedure as proposed in Arouna (2003) for the same set of model parameters. We replicated Table 6 in Capriotti (2008) using our method and compared the resulting variance ratios. As it turns out, our method yields average variance ratios that are 10% and 95% higher than the values reported by Capriotti (2008) and Arouna (2003), respectively.

Finally, in Table 7, Capriotti provides the results for a partial average Asian call option, as previously investigated in Su and Fu (2000). For detailed definitions of the models and parameters utilized in the simulations, we refer to Section 5 in Capriotti (2008). We implemented this particular model using our method. On average, our approach yielded variance ratios that were 10% and 50% higher than the values reported by Capriotti (2008) and Su and Fu (2000), respectively.

1.4.3. Multiple rare events

In Glasserman and Wang (1997), the authors emphasize that rare events often consist of unions of meaningful events that represent different ways in which the rare event can occur. In this context, we aim to examine an example where the rare event is formed by the intersection of two rare events. We will also discuss the case of the union of rare events later on. An illustrative example is provided by knock-out call options, which exhibit a classical scenario where the payoff is discontinuous with respect to the asset price trajectory. In this case, two potentially rare events can arise: (1) the arithmetic average $\bar{X}_t = \int_0^t \langle w, X_s \rangle \, \mathrm{d}s$ must be above the strike at the terminal time, and (2) the option must not be knocked out.

Consider an asset price X that follows a classical Black–Scholes model, characterized by the SDE $\mathrm{d}X_t = rX_t \, \mathrm{d}t + \sigma X_t \, \mathrm{d}B_t$, with an initial value of $X_0 = x$, where B denotes a Brownian motion. In contrast to the previous subsection, we introduce knock-out barriers L, U that satisfy $0 < L < X_0 < K < U$. The option is considered knocked out if the arithmetic average $\bar{X}_t$ breaches either of the two barriers at any given point in time before or at maturity. In our example, there is a delicate balance which needs to be achieved between giving the asset a positive drift such that $\bar{X}_1 > K$ with sufficiently high probability, and making sure that the option is not knocked out.

In Figure 1.2 we provide a graphical representation of the learning process of the neural network. On a fixed dataset, we calculate the probability of the arithmetic average ending

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

up above the strike K, the probability of it remaining between the knock-out barriers at all times, as well as the variance ratio after each epoch that the neural network was trained. Table 1.2 provides a comprehensive overview of the variance ratios corresponding to different values of strikes K and upper knock-out barriers U.

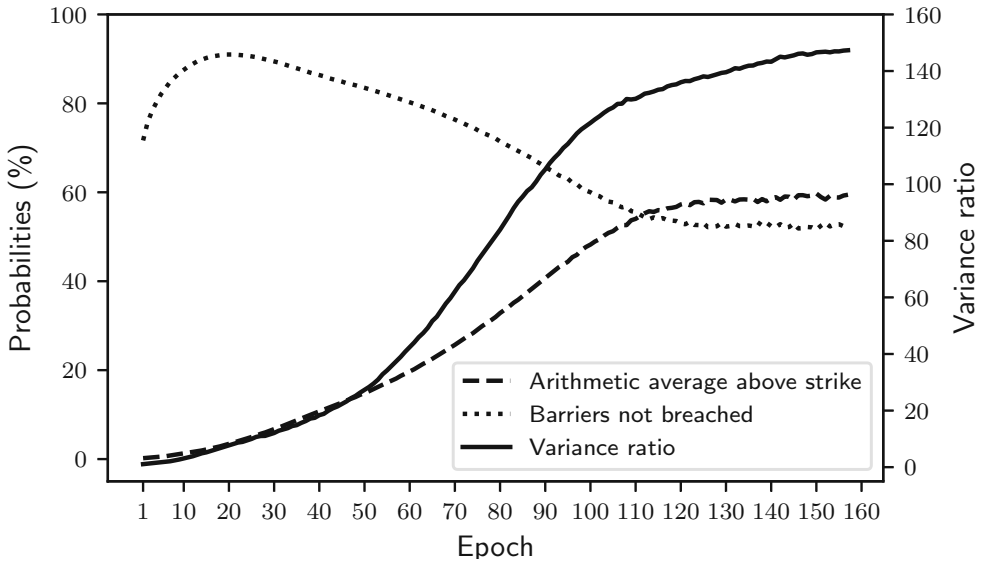

Figure 1.2. *A visual representation of the learning process of the neural network.*

Figure 1.2 highlights an interesting observation: increasing the variance ratio does not simply result from an indiscriminate rise in the probabilities of both rare events occurring. Instead, it becomes evident that a delicate balance between the occurrence of both rare events is crucial to increase the variance ratio. As demonstrated in Figure 1.2, neural networks exhibit the capability to learn and navigate this balancing act. Table 1.2 shows again that the neural network-induced change of measure is able to reduce the variance to varying degrees. We note that compared to the example of the previous subsection, adding stratification does not yield such a dramatic increase in variance ratio, however the improvement is still notable in most cases.

The model which we studied in this subsection has also been explored in Glasserman et al. (1999). In their paper, the authors report in Table 5.2 variance ratios for different values of the volatility, the strike as well as the knock-out barrier U (setting the lower knock-out barrier L to zero). In contrast to our model, the knock-out occurs in case the asset price breaches the knock-out barrier U at terminal time, i.e. in case $X_1 > U$. We replicated their model using our methodology and compared the achieved variance ratios. Our method on average achieved 20% higher variance ratios for the case of importance

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 1.2. *Variance ratios for different strikes K, volatilities σ and knock-out barriers U.*

Parameters			Importance Sampling			IS and Stratification		
K	σ	U	Mean	Std. err.	VR	Mean	Std. err.	VR
60	0.2	70	0.763	0.013 (1.70%)	7	0.779	0.012 (1.54%)	7
		80	12.605	0.067 (0.53%)	10	12.588	0.049 (0.39%)	18
		90	22.607	0.078 (0.35%)	18	22.673	0.032 (0.14%)	112
	0.3	70	0.1826	0.0082 (4.49%)	3	0.1816	0.0080 (4.41%)	3
		80	13.65	0.12 (0.88%)	4	13.60	0.11 (0.81%)	4
		90	42.86	0.22 (0.51%)	5	42.89	0.16 (0.37%)	9
70	0.2	80	0.000775	0.000041 (5.29%)	356	0.000760	0.000041 (5.39%)	357
		90	0.1917	0.0018 (0.94%)	144	0.1921	0.0016 (0.83%)	189
		100	0.6473	0.0035 (0.54%)	203	0.6449	0.0021 (0.33%)	537
	0.3	80	0.00070	0.00014 (20%)	36	0.00068	0.00014 (20.59%)	37
		90	0.724	0.011 (1.52%)	17	0.733	0.010 (1.36%)	18
		100	4.513	0.034 (0.75%)	18	4.507	0.029 (0.64%)	25

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes is 2. Other model parameters are $X_0 = 50$, $r = 0.05$ and $L = 40$.

sampling without stratification. However, when incorporating stratified sampling, our method on average achieved variance ratios that were 10% lower compared to those reported in Glasserman et al. (1999, Table 5.2). Note that the setting of Glasserman et al. (1999) is discrete in time, and that the authors consider asymptotically optimal drift adjustments. These findings suggest that there might be ample scope to further investigate optimal neural-network induced stratification for continuous-time models.

Let us now revisit the method proposed by Capriotti (2008). In Section 5 of his work, the author presents an example in the form of a European straddle: $F(X) = (X_1 - K)^+ + (K - X_1)^+$. Capriotti (2008) argues that in this case, the optimal sampling density would need to be bi-modal, a property that cannot be effectively captured by a normal distribution. As we attempted to implement this example, it became evident that the neural network struggled to determine the appropriate drift direction. This particular instance highlights the challenges associated with relying solely on drift adjustments for variance reduction. It serves as an example where the rare event can be characterized as the union of two events, shedding light on the limitations of such an approach.

1.4.4. Dynamic correlation

The generality of our approach builds on the decomposition $[M] = \int \pi(s)\, \mu(\mathrm{d}s)$. While Subsection 1.4.2 deviates from the conventional Brownian setting where $\mu = \lambda$, we also aim to present an example that diverges from the typical scenario examined in the existing

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

literature, where $\pi \equiv \mathrm{id}$, representing the identity matrix. To this end, we consider a Heston model with a dynamic variance-covariance matrix.

We assume that the price process X follows the dynamics given by the SDE $\mathrm{d}X_t = rX_t\,\mathrm{d}t + \sqrt{V_t}X_t\,\mathrm{d}B_t$. The instantaneous variance V follows CIR-type dynamics described by $\mathrm{d}V_t = \kappa(\theta - V_t)\,\mathrm{d}t + \xi\sqrt{V_t}\,\mathrm{d}W_t$. B and W are correlated Brownian motions related through $\mathrm{d}[B,W]_t = \rho(t)\,\mathrm{d}t$, where the correlation function takes the form $\rho(x) = \bar{\rho} + \bar{\rho}A\sin(2\pi f x)$. In other words, we deviate from the constant correlations regime by means of the multiple of a sine wave with amplitude A and frequency f. We present the results for various combinations of amplitude and frequency choices in Table 1.3 below.

Table 1.3. *Variance ratios for different values of amplitude A frequency f.*

Param.		**Importance Sampling**			**IS and Stratification**		
A	f	Mean	Std. err.	Var. ratio	Mean	Std. err.	Var. ratio
0	0	2.2145	0.0085 (0.38%)	171	2.2308	0.0063 (0.28%)	311
0.2	1	1.9378	0.0076 (0.39%)	178	1.9517	0.0058 (0.30%)	304
	2	2.0807	0.0082 (0.39%)	171	2.0938	0.0062 (0.30%)	297
	4	2.1498	0.0085 (0.40%)	165	2.1651	0.0065 (0.30%)	283
0.5	1	1.5544	0.0062 (0.40%)	203	1.5666	0.0048 (0.31%)	338
	2	1.8926	0.0077 (0.41%)	173	1.9037	0.0059 (0.31%)	289
	4	2.0553	0.0081 (0.38%)	168	2.0691	0.0062 (0.30%)	289
1	1	1.0147	0.0041 (0.39%)	268	1.0239	0.0032 (0.31%)	443
	2	1.6117	0.0064 (0.40%)	202	1.6230	0.0047 (0.29%)	369
	4	1.9048	0.0078 (0.41%)	165	1.9160	0.0060 (0.31%)	277

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes is 5. Other model parameters are $X_0 = 50$, $r = 0.05$, $V_0 = 0.04$, $\kappa = 2$, $\theta = 0.09$, $\xi = 0.2$, $\bar{\rho} = -0.5$ and $K = 70$.

1.4.5. Basket option

So far, we have presented results in scenarios with low dimensions. However, the multi-dimensional formulation of our setting suggests investigating whether we can achieve satisfactory levels of variance reduction for higher-dimensional models. Inspired by Jourdain and Lelong (2009), we study a multi-dimensional Black–Scholes model.

Consider the d-dimensional asset price X governed by the SDE $\mathrm{d}X_t = r \odot X_t\,\mathrm{d}t + X_t \odot \mathrm{d}M_t$, where $M_t = \Sigma B_t$ represents a d-dimensional standard Brownian motion B_t with variance-covariance matrix $\Sigma\Sigma^\top$. We sample the initial value X_0 of X uniformly from the range of 10 to 200. Moreover, we sample the vector r of appreciation rates and the vector σ of volatilities uniformly between 1% and 9% as well as 10% and 30%, respectively. The

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

weight vector w is then computed as $w_i = r_i/\sigma_i^2$ and further normalized to sum to 1.

To define the matrix $\Sigma\Sigma^\mathsf{T}$, it is necessary to specify the correlation matrix. In order to ensure a valid correlation matrix that remains positive definite even in high dimensions, we adopt the approach proposed by Davies and Higham (2000). Firstly, we sample a d-dimensional vector y uniformly between 0 and 1. We then re-scale the vector y such that the sum of its elements equals the dimension d. The algorithm proposed in Davies and Higham (2000) then generates a valid correlation matrix, whose eigenvalues correspond to the values in the re-scaled vector y. Finally, we still need to specify the strike. To this end, we sample 10^4 observations of the arithmetic average $\bar{X}$ at maturity, and then choose the strike K to approximately be above the 90th percentile of the distribution of $\bar{X}_1$. Note that the choice of K is highly dependent on the previously sampled parameters.

Table 1.4. *Variance ratios for different dimensions d.*

Parameters		**Importance Sampling**				
d	K	Mean	Std. err.	Var. ratio	$\mathbb{P}[F(X) > 0]$	$\mathbb{Q}[F(X) > 0]$
10	88	3.7557	0.0118 (0.31 %)	62	3.24 %	70.83 %
20	115	1.3252	0.0049 (0.37 %)	124	1.22 %	65.42 %
50	126	6.4457	0.0189 (2.93 %)	28	6.96 %	72.36 %
100	106	1.6995	0.0056 (0.33 %)	54	3.36 %	70.69 %
200	112	2.1373	0.0067 (0.31 %)	36	5.17 %	72.04 %
500	110	1.1352	0.0042 (0.37 %)	30	4.46 %	74.99 %
1,000	110	2.327	0.011 (0.47 %)	6	10.87 %	84.72 %

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes corresponds to the dimension d. $\mathbb{P}[F(X) > 0]$ represents the proportion of trajectories in the test dataset where the payoff is positive, without incorporating a drift adjustment. $\mathbb{Q}[F(X) > 0]$ denotes the proportion of trajectories in the test dataset where the payoff is positive under the drift adjustment.

Table 1.4 presents variance ratios obtained for various dimensions d ranging from $d = 10$ up to $d = 1{,}000$. Moreover, we also compared our method to the approach presented in Jourdain and Lelong (2009). In their study, the authors considered the 40-dimensional case, and all volatilities, appreciation rates, and weights were chosen uniformly across all assets in the basket. We refer to Section 3 in Jourdain and Lelong (2009) for further details about the model as well as model parameters in their Table 1. As it turns out, our method achieves variance ratios that are, on average, comparable to those reported by Jourdain and Lelong (2009). It is important to note that the strikes which were chosen are relatively close to the initial value. In previous examples, we can observe that the obtained variance ratios tend to grow as the strike is increased. In contrast to Jourdain and Lelong (2009), we present in Table 1.4 results for dimension up to $d = 1{,}000$, which we believe is a distinctive aspect worth highlighting.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Conclusions

We presented a method that uses feedforward neural networks for the purpose of reducing the variance of Monte Carlo estimators. To this end, we studied the class of Gaussian measures which are induced by vector-valued continuous local martingales with deterministic covariation. Building on the theory of vector stochastic calculus, we identified the Cameron–Martin spaces of those measures, and proved universal approximation theorems that establish, up to an isometry, topological density of feedforward neural networks in these spaces. We then applied our results to a classical importance sampling approach which seeks for an optimal drift adjustment of the processes which are driving the asset prices. Finally, we presented the results of a numerical study, which clearly indicate the potential of this approach.

Let us also remark that our approach comes with several challenges. In principle, one needs to train separate feedforward networks for different models and model parameters. In light of Remark 1.41, one could train a feedforward network to minimize a weighted average standard error over several models or model parameters. Complex, high-dimensional models might call for the need of using complex neural network architectures in order to achieve a sufficient variance reduction, which might lead to a considerable computational effort for training the feedforward networks. On the other hand, the competing approaches Guasoni and Robertson (2008); Robertson (2010) involve having to solve a potentially complex, high-dimensional variational problem, whose solution might involve a numerical procedure which might induce a considerable computational effort, too. Finally, while Theorem 1.38 and the simulations of Section 1.4 show that one can obtain a sufficient variance reduction with shallow feedforward networks, the model-dependent choice of optimal architecture has not been discussed at all, which highlights the potential for a further improvement of this method.

Outlook on further research

Throughout this chapter, we assumed for the process M to be a continuous local martingale with deterministic covariation, such that it is a Gaussian process and induces a Gaussian measure on path space. Clearly, there are Gaussian processes which cannot be local martingales, e.g. fractional Brownian motion with Hurst index $\neq 1/2$. In line with Remark 1.15, Section 1.1 can be extended to the study of multivariate Volterra type Gaussian processes of the form $\tilde{M}_t = \int_0^T k(t,s)\,\mathrm{d}M_s$ with a matrix-valued kernel k. While Section 1.3 makes use of the semimartingale property of M by applying Girsanov's theorem and studying convergence of stochastic exponentials, the Cameron–Martin theorem (see Theorem 1.46) can still be applied to the Gaussian measure that is induced by $\tilde{M}$ on path space. These considerations in particular motivate the study of a refined class of multivariate (fractional) stochastic volatility models, their small-time asymptotics as well as importance sampling methods for the numerical evaluation of derivatives for

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

these models, which is subject to a follow-up work.

In Section 1.3, we required $F(X)$ to be $\mathcal{F}_T$-measurable and L^p-integrable for some $p > 2$. However, the properties that we imposed on the process X were rather weak. In particular, Theorem 1.38 only considered $F(X)$ as a random variable, where we used the SDE for X only when performing a measure change and applying Girsanov's theorem in order to understand the semimartingale decomposition of X under a new sampling measure. Therefore, the methods from Section 1.3 should extend to the case where X is the solution to a McKean–Vlasov SDE, provided that we understand how the dynamics of the process change under a change of measure. We leave it to a follow-up work to combine our methods with ideas from dos Reis et al. (2023), which should lead to a tractable importance sampling framework for the evaluation of derivatives on solutions to McKean–Vlasov SDEs under weaker assumptions then those imposed on dos Reis et al. (2023).

The setting of this chapter does naturally apply to the evaluation of European options and asset price processes with continuous paths. More generally, reducing the standard error of Monte Carlo estimators with neural networks when pricing American options based on the popular algorithm proposed by Longstaff and Schwartz (cf. Clément et al. (2002); Longstaff and Schwartz (2001)) and models with jumps, very much in the spirit of Genin and Tankov (2020) as well as Kawai (2009), provides another interesting challenge that is reserved for follow-up work.

Finally, the measure changes which we studied in Section 1.3 were induced by density processes of the form $\mathcal{E}(f^\mathsf{T} \bullet M)$, where $f \in \Lambda^2$ is a deterministic function. The reason why we did not consider the more general class of processes $U \in L^2(M)$ for which $\mathcal{E}(U^\mathsf{T} \bullet M)$ is a martingale is twofold. While the proof of Theorem 1.38 would become more involved, one would need to use neural network architectures which are more complex then the ones which we discussed in Section 1.2. For this reason, we argue that the problem of considering deterministic functions $f \in \Lambda^2$ provides a tractable, numerically efficient method to reduce the variance in Monte Carlo simulations, and reserve the extension to processes $U \in L^2(M)$ and their approximation with neural networks for future work.

1.5. Supplementary material

1.5.1. Gaussian measures

In this subsection, we collect for the readers' convenience some classical definitions and results about Gaussian measures, for which we mostly rely on the excellent monographs Bogachev (1998); Lifshits (2012); Stroock (2011). Let $(E, \|\cdot\|_E)$ denote a real separable Banach space, γ a Borel probability measure on E and $M = (M_t)_{t\in[0,T]}$ an $\mathbb{R}^d$-valued process on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given $h \in E$, we further denote by γ_h the measure on E that is induced by the translation $E \ni x \mapsto x + h$.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Definition 1.42. The measure γ is called Gaussian, if each $f \in E^*$ induces a Gaussian distribution on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. The process $M = (M_t)_{t\in[0,T]}$ is called Gaussian, if $(M_{t_i})_{i=1}^n$ is jointly Gaussian for each $n \in \mathbb{N}$ and $0 \le t_1 < t_2 < \ldots < t_n \le T$.

A Gaussian measure γ is centered, if each $f \in E^*$ induces a centered Gaussian distribution. Similarly, a Gaussian process $M = (M_t)_{t\in[0,T]}$ is centered, if $(M_{t_i})_{i=1}^n$ is jointly centered Gaussian for each $n \in \mathbb{N}$ and $0 \le t_1 < t_2 < \ldots < t_n \le T$. Since Section 1.1 only considers centered Gaussian processes and measures, we will from now on restrict to this special case.

In the context of Definition 1.42, we have the natural embedding $j\colon E^* \to E^*_\gamma$, where E^*_γ denotes the reproducing kernel Hilbert space of γ, which is defined as the closure of E^* in $L^2(\gamma)$. We further define the covariance operator of γ by the map

$$R_\gamma\colon E^* \to (E^*)'\colon f \mapsto \left(g \mapsto \int_E f(x)g(x)\gamma(\mathrm{d}x)\right),$$

and implicitly consider its extension to E^*_γ, i.e. $R_\gamma\colon E^*_\gamma \to (E^*)'$.

Given $f \in E^*_\gamma$, note that $R_\gamma(f)\colon E^* \to \mathbb{R}$ is a linear operator. If we endow E^* with the Mackey topology, then Bogachev (1998, Lemma 3.2.1) shows that $R_\gamma(f)$ is continuous. Mackey's theorem (cf. Bogachev (1998, Theorem A 1.1)) yields the existence of $x_f \in E$, such that $R_\gamma(f)(g) = g(x_f)$ for each $g \in E^*$. We then also denote by R_γ the map $E^*_\gamma \ni f \mapsto x_f$.

Definition 1.43. Given a centered Gaussian measure γ on E, the Cameron–Martin space $H(\gamma)$ of γ is defined as the range of R_γ in E, i.e. $H(\gamma) \coloneqq R_\gamma(E^*_\gamma) \subset E$. We equip $H(\gamma)$ with the inner product

$$\langle h, k\rangle_{H(\gamma)} \coloneqq \langle \hat{h}, \hat{k}\rangle_{L^2(\gamma)} = \int_E \hat{h}(x)\hat{k}(x)\gamma(\mathrm{d}x), \quad h, k \in H(\gamma),$$

where $h = R_\gamma(\hat{h})$ and $k = R_\gamma(\hat{k})$ for some $\hat{h}, \hat{k} \in E^*_\gamma$.

The space $(H(\gamma), \langle\cdot,\cdot\rangle_{H(\gamma)})$ is a real separable Hilbert space that is continuously embedded into E (cf. Bogachev (1998, Proposition 2.4.6 and Theorem 3.2.7)). Moreover, Bogachev (1998, Theorem 2.4.7) shows that $H(\gamma)$ is of γ-measure zero, whenever E^*_γ is infinite dimensional.

Remark 1.44. Given a centered Gaussian measure γ on E, the topological support of γ is defined as the smallest closed subset $S \subset E$ with $\gamma(E \setminus S) = 0$, and is given by $\overline{H(\gamma)}$, where the closure is taken in E (cf. Bogachev (1998, Theorem 3.6.1)). We call γ nondegenerate, if $\overline{H(\gamma)} = E$, or equivalently, if $H(\gamma)$ is densely embedded into E. If $\overline{H(\gamma)}$ is a strict subspace of E, then we call γ degenerate.

Remark 1.45. If γ and $\tilde{\gamma}$ are two centered Gaussian measures on E with $H(\gamma) = H(\tilde{\gamma})$ and $\|\cdot\|_{H(\gamma)} = \|\cdot\|_{H(\tilde{\gamma})}$, then γ and $\tilde{\gamma}$ coincide (cf. Bogachev (1998, Corollary 3.2.6)). Moreover,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

if E is continuously and linearly embedded into another real separable Banach space $\tilde{E}$ with embedding i and induced Gaussian measure $\nu = \gamma \circ i^{-1}$, then $\tilde{E} \supset H(\nu) = i(H(\gamma))$ (cf. Bogachev (1998, Lemma 3.2.2)).

The Cameron–Martin space has another useful characterization, which is stated in the following theorem (cf. Bogachev (1998, Theorem 2.4.5), Cameron and Martin (1944, Theorem 1)).

Theorem 1.46. *Given a centered Gaussian measure γ on E and $h \in E$, the measures γ and γ_h are equivalent precisely when $h \in H(\gamma)$, and singular otherwise. In particular,*

$$H(\gamma) = R_\gamma(E^*_\gamma) = \{h \in E : \gamma_h \sim \gamma\}.$$

Whenever γ is Gaussian, the measure γ_h is Gaussian for each $h \in E$ (cf. Bogachev (1998, Lemma 2.2.2)). Consequently, Theorem 1.46 characterizes a set of Gaussian measures which are equivalent to γ. The following theorem (cf. Bogachev (1998, Theorem 2.7.2)) is another central result, which in particular implies that γ_h and γ are singular whenever $h \in E \setminus H(\gamma)$.

Theorem 1.47. *Any two Gaussian measures on E are either equivalent or mutually singular.*

In order to quantify the (exponential) decline of the probability of certain tail events, the following result is often times useful (cf. Bogachev (1998, Corollary 4.9.3)).

Proposition 1.48. *Let γ denote a centered Gaussian measure on E. Moreover, for $\varepsilon > 0$, let γ_ε denote the pushforward measure of γ under the map $E \ni f \mapsto \sqrt{\varepsilon} f$. Then, $(\gamma_\varepsilon)_{\varepsilon>0}$ satisfies the large deviation principle with rate function $I_\gamma \colon E \to \overline{\mathbb{R}}_+$, where*

$$I_\gamma(f) = \begin{cases} \frac{1}{2}\|f\|^2_{H(\gamma)} & \text{for } f \in H(\gamma), \\ \infty & \text{otherwise.} \end{cases}$$

In other words, for each $F \in \mathcal{B}_E$,

$$-\inf_{f \in F^\circ} I_\gamma(f) \le \liminf_{\varepsilon \searrow 0} \varepsilon \log \gamma_\varepsilon(F) \le \limsup_{\varepsilon \searrow 0} \varepsilon \log \gamma_\varepsilon(F) \le -\inf_{f \in \bar{F}} I_\gamma(f).$$

Before we finish this section, we state a result that allows us in many cases to obtain a tractable representation of $(H(\gamma), \langle \cdot, \cdot \rangle_{H(\gamma)})$ (cf. Bogachev (1998, Section 3.3) and Lifshits (2012, Section 4.2)).

Theorem 1.49. *Given a centered Gaussian measure γ on E, assume that there exists a Hilbert space $\tilde{H}$ and a continuous linear operator $J \colon \tilde{H} \to E$ such that R_γ admits the*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

factorization $R_\gamma = J \circ J^*$, *where* $J^*\colon E^* \to \tilde{H}^* \cong \tilde{H}$ *denotes the adjoint of* J. *Then* $H(\gamma)$ *coincides with* $J(\tilde{H})$. *If* J *is moreover injective, then*

$$\langle f, g\rangle_{H(\gamma)} = \langle J^{-1}(f), J^{-1}(g)\rangle_{\tilde{H}}, \quad f, g \in H(\gamma).$$

1.5.2. Proofs

Proof of Lemma 1.10(a). The proof of Cherny and Shiryaev (2002, Lemma 3.2) reveals that $\|\cdot\|_{\Lambda^2}$ satisfies the triangle inequality, which shows that Λ^2 is a real vector space. In order to see that $\langle\cdot,\cdot\rangle_{\Lambda^2}$ is an inner product on Λ^2, note that, by construction, $\langle\cdot,\cdot\rangle_{\Lambda^2}$ is symmetric and linear in both arguments, and recall that π is positive semidefinite μ-almost everywhere, hence $f^\mathsf{T}\pi f \geq 0$ μ-almost everywhere and therefore $\int_0^T f^\mathsf{T}(s)\pi(s)f(s)\,\mu(\mathrm{d}s) \geq 0$ for each measurable $f\colon [0,T] \to \mathbb{R}^d$. If $\langle f, f\rangle_{\Lambda^2} = 0$ for some $f \in \Lambda^2$, then $f^\mathsf{T}\pi f = 0$ μ-almost everywhere, hence $f \sim 0$, which implies that $\langle\cdot,\cdot\rangle_{\Lambda^2}$ is positive definite and therefore an inner product on Λ^2.

Completeness of (Λ^2, ϱ_2), where ϱ_2 denotes the translation invariant metric induced by $\|\cdot\|_{\Lambda^2}$, follows from Jacod (1979, Lemme 4.29), and separability can be argued by adapting the proofs of Billingsley (2012, Theorem 19.2) and Cherny and Shiryaev (2002, Lemma 3.2). We conclude that $(\Lambda^2, \langle\cdot,\cdot\rangle_{\Lambda^2})$ is a real separable Hilbert space. □

Lemma 1.10(b) is a direct consequence of the Fréchet–Riesz representation theorem, since we know by Lemma 1.10(a) that Λ^2 is a Hilbert space.

Proof of Lemma 1.10(c). $\Lambda^{2,0}$ is clearly a real vector space. Given $f \in \Lambda^{2,0}$ and $i, j \in \{1, 2, \ldots, d\}$, by a version of the Kunita–Watanabe inequality for Lebesgue–Stieltjes integrals (cf. Schmock (2024, Lemma 5.91)),

$$\begin{aligned}\Big(\Big|\int_0^T f_i(s)f_j(s)\mu_{i,j}(\mathrm{d}s)\Big|\Big)^2 &\leq \Big(\int_0^T |f_i(s)f_j(s)||\mu_{i,j}|(\mathrm{d}s)\Big)^2 \\ &\leq \int_0^T f_i^2(s)\mu_{i,i}(\mathrm{d}s)\int_0^T f_j^2(s)\mu_{j,j}(\mathrm{d}s) < \infty,\end{aligned} \tag{1.12}$$

hence $\Lambda^{2,0} \subset \Lambda^2$, and $(\Lambda^{2,0}, \langle\cdot,\cdot\rangle_{\Lambda^2})$ is therefore an inner product space.

The fact that $\Lambda^{2,0}$ is dense in Λ^2 has been shown in Jacod (1979, Lemme 4.29), which also implies the separability of $\Lambda^{2,0}$. From Lemma 1.10(a) we further know that $(\Lambda^2, \langle\cdot,\cdot\rangle_{\Lambda^2})$ is a Hilbert space and in particular complete. This shows 1.10(c). □

Proof of Lemma 1.10(d). The continuity of the embedding follows from the inequalities (1.12) and (1.15). The remaining assertion follows from a multivariate version of Kallenberg (2021, Lemma 1.37). □

Proof of Lemma 1.10(e). Let $(\tilde{\pi}, \tilde{\mu})$ be another pair that satisfies the representation (1.1), and let $f \in \Lambda^{2,0}$. Then $\mathrm{d}\mu_{i,j}/\mathrm{d}\mu = \pi_{i,j}$ as well as $\mathrm{d}\mu_{i,j}/\mathrm{d}\tilde{\mu} = \tilde{\pi}_{i,j}$ for all $i, j \in \{1, 2, \ldots, d\}$,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

hence, using (1.12),

$$\int_0^T f^\mathsf{T}(s)\pi(s)f(s)\,\mu(\mathrm{d}s) = \sum_{i,j=1}^d \int_0^T f_i(s)f_j(s)\mu_{i,j}(\mathrm{d}s) = \int_0^T f^\mathsf{T}(s)\tilde{\pi}(s)f(s)\tilde{\mu}(\mathrm{d}s),$$

which extends to all $f \in \Lambda^2$ using the density of $\Lambda^{2,0}$ in Λ^2, and we see that $\|\cdot\|_{\Lambda^2}$ does not depend on the specific choice of (π,μ) satisfying (1.1). Moreover, for $f,g\colon [0,T]\to\mathbb{R}^d$ measurable, $(f-g)^\mathsf{T}\pi(f-g)=0$ μ-almost everywhere holds precisely when $\|f-g\|_{\Lambda^2}=0$ which is equivalent to $(f-g)^\mathsf{T}\tilde{\pi}(f-g)=0$ $\tilde{\mu}$-almost everywhere. □

Proof of Proposition 1.18(a). By a variant of the Cauchy–Schwarz inequality, for $h\in H$, $i\in\{1,2,\dots,d\}$ and $t\in[0,T]$, it holds that (Cherny and Shiryaev, 2002, Lemma 4.17)

$$\int_0^t \left|\sum_{j=1}^d \pi_{i,j}(s)f_{h,j}(s)\right| \mu(\mathrm{d}s) \le \sqrt{\mu_{i,i}([0,t])}\|f_h\|_{\Lambda^2} \le \sqrt{\mu([0,T])}\|h\|_H, \tag{1.13}$$

which shows that the integral in (1.2) is well defined. □

Proof of Proposition 1.18(b). H is clearly a real vector space and $\langle\cdot,\cdot\rangle_H$ is symmetric and linear in both arguments. In order to show that $\langle\cdot,\cdot\rangle_H$ is positive definite, note that $\langle h,h\rangle_H = \|f_h\|_{\Lambda^2}^2 \ge 0$ for each $h\in H$. If $h\in H$ satisfies $\langle h,h\rangle_H=0$, then $f_h^\mathsf{T}\pi f_h=0$ μ-almost everywhere, hence $f_h\sim 0$. An application of inequality (1.13) shows that $(\pi f_h)_i=0$ μ-almost everywhere for each $i\in\{1,2,\dots,d\}$, hence $h=0$. $(H,\langle\cdot,\cdot\rangle_H)$ is therefore an inner product space.

We obtain a norm $\|\cdot\|_H$ on H by setting $\|h\|_H \coloneqq \sqrt{\langle h,h\rangle}_H$ and thus also a metric ϱ_H on H by setting $\varrho_H(f,g)\coloneqq\|f-g\|_H$. In order to see that (H,ϱ_H) is complete, let $(h_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in H. Then $(f_{h_n})_{n\in\mathbb{N}}$ is a Cauchy sequence in Λ^2. From Lemma 1.10(a) we know that Λ^2 is complete. Consequently, there exists some $f\in\Lambda^2$ such that $f_{h_n}\to f$ in Λ^2. If we set $h=J(f)$, then $h\in H$ and $h_n\to h$ in H.

Finally, in order to see that H is separable, note first that Λ^2 is separable by Lemma 1.10(a). But this already shows that H is separable as well, because a countable dense subset of H is given by $\{h\in H\colon f_h\in B\}$, where B is a countable dense subset of Λ^2. □

Proof of Proposition 1.18(c). By construction, $J\colon\Lambda^2\to H$ is a linear isometry. Since $\Lambda^{2,0}$ is a linear subspace of Λ^2 by Lemma 1.10(c), we see that $(H^0,\langle\cdot,\cdot\rangle_H)$ is an inner product subspace of H. If $(h_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in H^0, then $(f_{h_n})_{n\in\mathbb{N}}$ is a Cauchy sequence in $\Lambda^{2,0}$. By Lemma 1.10(c) there exists an $f\in\Lambda^2$ such that $f_{h_n}\to f$ as $n\to\infty$. Denoting $h=J(f)\in H$, it follows that $h_n\to h$ in H, as $n\to\infty$. □

Proposition 1.18(d) is a direct consequence of the Fréchet–Riesz representation theorem, since we know by Proposition 1.18(b) that H is a Hilbert space.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Remark 1.50. For the proof of Proposition 1.18(e), we use a multivariate version of the Riesz–Markov–Kakutani representation theorem, which states that every $F \in (C([0,T];\mathbb{R}^d))^*$ can be identified with an $\mathbb{R}^d$-valued function $\nu = (\nu_1, \nu_2, \dots, \nu_d)^\mathsf{T}$ on $\mathcal{B}_{[0,T]}$, where every entry is a signed Borel measure of finite total variation, such that

$$F(f) = \sum_{j=1}^{d} \int_0^T f_j(s)\nu_j(\mathrm{d}s) =: \int_0^T f^\mathsf{T}(s)\nu(\mathrm{d}s), \quad f \in C([0,T];\mathbb{R}^d).$$

Given $f\colon [0,T] \to \mathbb{R}^{n\times d}$ such that $(f_{i,\cdot})^\mathsf{T} \in C([0,T];\mathbb{R}^d)$ for each $i \in \{1,2,\dots,n\}$, we write

$$\int_0^T f(s)\nu(\mathrm{d}s) = \Big(\int_0^T f_{1,\cdot}(s)\nu(\mathrm{d}s), \dots, \int_0^T f_{n,\cdot}(s)\nu(\mathrm{d}s) \Big)^\mathsf{T}.$$

For generalizations to infinite-dimensional domains and image spaces, see Gowurin (1936); Singer (1957).

Proof of Proposition 1.18(e). We argue in line with Lifshits (2012, Example 4.4) and use Theorem 1.49. First, we note that every $h \in H$ is continuous and satisfies $h(0) = 0$, hence $H \subset E$. Let us consider J as a linear operator onto E, i.e. $J\colon \Lambda^2 \to E$, which is continuous due to (1.13). In the context of Remark 1.50, E^* is given as the quotient space, where we identify those $\nu \in (C([0,T];\mathbb{R}^d))^*$ that annihilate E.

For $f \sim \sigma$ and $g \sim \nu$ in E^*,

$$\begin{aligned}
R_\gamma(f)(g) &= \int_E f(x)g(x)\gamma_M(\mathrm{d}x) = \mathbb{E}[f(M)g(M)] = \mathbb{E}\Big[\int_0^T M_s^\mathsf{T}\sigma(\mathrm{d}s) \int_0^T M_s^\mathsf{T}\nu(\mathrm{d}s)\Big] \\
&= \sum_{i,j=1}^{d} \int_0^T \int_0^T \mathbb{E}[M_s^i M_t^j]\sigma_i(\mathrm{d}s)\nu_j(\mathrm{d}t) = \sum_{i,j=1}^{d} \int_0^T \int_0^T [M]_{s\wedge t}^{i,j}\sigma_i(\mathrm{d}s)\nu_j(\mathrm{d}t) \\
&= \sum_{j=1}^{d} \int_0^T \sum_{i=1}^{d} \int_0^T [M]_{s\wedge t}^{i,j}\sigma_i(\mathrm{d}s)\nu_j(\mathrm{d}t) = \sum_{j=1}^{d} \int_0^T \Big(\int_0^T [M]_{s\wedge t}\sigma(\mathrm{d}s) \Big)_j \nu_j(\mathrm{d}t) \\
&= \int_0^T \Big(\int_0^T [M]_{s\wedge t}\sigma(\mathrm{d}s) \Big)^\mathsf{T} \nu(\mathrm{d}t) = \Big(g, \int_0^T [M]_{s\wedge\cdot}\sigma(\mathrm{d}s) \Big).
\end{aligned}$$

We can therefore identify $R_\gamma(f)$ with $\int_0^T [M]_{s\wedge\cdot}\sigma(\mathrm{d}s)$.

Next, let us find the adjoint of J. Given $f \in \Lambda^2$ and $g \sim \nu$ in E^*,

$$(g, J(f)) = \sum_{i=1}^{d} \int_0^T J_i(f)(t)\nu_i(\mathrm{d}t) = \sum_{i=1}^{d} \int_0^T \int_0^t \pi_{i,\cdot}(s)f(s)\,\mu(\mathrm{d}s)\nu_i(\mathrm{d}t)$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

$$
\begin{aligned}
&= \sum_{i=1}^{d} \int_0^T \int_0^T \mathbb{1}_{[0,t]}(s)\pi_{i,\cdot}(s)f(s)\,\mu(\mathrm{d}s)\nu_i(\mathrm{d}t) \\
&= \sum_{i=1}^{d} \int_0^T \int_0^T \mathbb{1}_{[0,t]}(s)\nu_i(\mathrm{d}t)\pi_{i,\cdot}(s)f(s)\,\mu(\mathrm{d}s) \\
&= \sum_{i=1}^{d} \int_0^T \nu_i([s,T])\pi_{i,\cdot}(s)f(s)\,\mu(\mathrm{d}s) \\
&= \int_0^T \nu([s,T])^{\mathsf{T}}\pi(s)f(s)\,\mu(\mathrm{d}s) = \langle \nu([\cdot,T]), f\rangle_{\Lambda^2},
\end{aligned}
$$

hence the adjoint $J^*\colon E^* \to (\Lambda^2)^* \cong \Lambda^2$ is given by $g \sim \nu \mapsto ([0,T] \ni s \mapsto \nu([s,T]))$.

Finally, the covariance operator admits for $g \sim \nu$ in E^* the factorization

$$
\begin{aligned}
R_\gamma(g)_i(t) &= \int_0^T [M]^{i,\cdot}_{s\wedge t}\nu(\mathrm{d}s) = \sum_{j=1}^{d} \int_0^T [M]^{i,j}_{s\wedge t}\nu_j(\mathrm{d}s) \\
&= \sum_{j=1}^{d} \int_0^T \int_0^{s\wedge t} \pi_{i,j}(w)\mu(\mathrm{d}w)\nu_j(\mathrm{d}s) \\
&= \sum_{j=1}^{d} \int_0^T \pi_{i,j}(w) \int_0^T \mathbb{1}_{[0,s\wedge t]}(w)\nu_j(\mathrm{d}s)\mu(\mathrm{d}w) \\
&= \sum_{j=1}^{d} \int_0^t \pi_{i,j}(w)\nu_j([w,T])\mu(\mathrm{d}w) \\
&= \int_0^t \pi_{i,\cdot}(w)\nu([w,T])\mu(\mathrm{d}w) \\
&= \int_0^t \pi_{i,\cdot}(w)J^*(g)(w)\mu(\mathrm{d}w) = (J\circ J^*)(g)_i(t),
\end{aligned}
$$

where $i \in \{1,2,\dots,d\}$ and $t \in [0,T]$.

Let us show that J is injective. Let $f_1, f_2 \in \Lambda^2$ be such that $J(f_1) = J(f_2)$, i.e. $\|J(f_1) - J(f_2)\|_\infty = 0$, which implies in particular for $g = f_1 - f_2$ that

$$
\int_A \pi(s)g(s)\,\mu(\mathrm{d}s) = 0 \in \mathbb{R}^d \tag{1.14}
$$

for all $A \in \mathcal{B}_{[0,T]}$ of the form $A = (s,t]$ for $s < t$ in $[0,T]$. Since the half-open intervals generate $\mathcal{B}_{[0,T]}$, Dynkin's theorem shows that (1.14) extends to all $A \in \mathcal{B}_{[0,T]}$.

We now show that $g \sim 0$, i.e. $g^{\mathsf{T}}\pi g = 0$ μ-almost everywhere. If this were not the case, then we would have, without loss of generality, $\mu(\{g^{\mathsf{T}}\pi g > 0\}) > 0$. We claim that $\{g^{\mathsf{T}}\pi g > 0\} \subset \{\pi g \neq 0\}$. To see this, pick $s \in [0,T]$ such that $g^{\mathsf{T}}(s)\pi(s)g(s) > 0$, and

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

assume that $\pi(s)g(s) = 0$. In other words, for each $i \in \{1, 2, \dots, d\}$, we would have $\sum_{j=1}^{d} \pi_{i,j}(s)g_j(s) = 0$. But this can't be the case, since we would then have

$$0 < g^{\mathsf{T}}(s)\pi(s)g(s) = \sum_{i=1}^{d} g_i(s)\Big(\sum_{j=1}^{d} \pi_{i,j}(s)g_j(s)\Big) = 0.$$

Now $\mu(\{g^{\mathsf{T}}\pi g > 0\}) > 0$ implies that $\mu(\{\pi g \neq 0\}) > 0$. Since $\{\pi g \neq 0\} = \bigcup_{i=1}^{d}\{(\pi g)_i \neq 0\}$, there exists $i \in \{1, 2, \dots, d\}$ such that $\mu(\{(\pi g)_i \neq 0\}) > 0$.

Without loss of generality, we may assume that $\mu(\{(\pi g)_i > 0\}) > 0$. Note that the set $A = \{(\pi g)_i > 0\}$ can be written as

$$A = \bigcup_{n\in\mathbb{N}} \{(\pi g)_i \geq \frac{1}{n}\} = \bigcup_{n\in\mathbb{N}} \left((\pi g)_i\right)^{-1}([\frac{1}{n}, \infty)),$$

where every $A_n := ((\pi g)_i)^{-1}([\frac{1}{n}, \infty))$ and therefore also A is $\mathcal{B}_{[0,T]}$-measurable. Now $\mu(A) > 0$ implies that $\mu(A_n) > 0$ for some $n \in \mathbb{N}$, hence $n \int_{A_n} (\pi g)_i(s)\, \mu(\mathrm{d}s) \geq \mu(A_n)$, which yields a contradiction to (1.14).

We may therefore conclude that $f_1 - f_2 = g \sim 0$ in Λ^2, which shows that the operator $J\colon \Lambda^2 \to E$ is injective. Theorem 1.49 now implies that the Cameron–Martin space of γ_M is given by $J(\Lambda^2) = H$. □

Proof of Theorem 1.24. The property of $H(D)$ being a dense subset of H follows from Convention 1.22 and the definition of the norm on H that is induced by the inner product $\langle\cdot,\cdot\rangle_H$. Being a dense subset of a separable metric space implies the remaining assertion of 1.24(a).

If D is also a linear subspace of Λ^2, then $H(D)$ is clearly an inner product space, whose completion is H by 1.24(a). We now follow a standard argument, a version of which can be found e.g. in Mercer (1986, Proposition 1). Since $H(D)$ is dense in H by 1.24(a) and H is separable due to Proposition 1.18(b), there exists a countable subset of $H(D)$ that is also dense in H. Upon applying the Gram–Schmidt process to this subset, one obtains a countable set of orthonormal vectors, which are in $H(D)$, whose linear span is dense in H.

We know from Proposition 1.18(e) that H is the Cameron–Martin space of γ_M. By standard theory for Gaussian measures we know that the topological support of γ_M then coincides with $\overline{H}$, where the closure is taken in E (see Remark 1.44). But since $H(D)$ is dense in H by 1.24(a), and the canonical injection from H to E is continuous by Bogachev (1998, Proposition 2.4.6), we have $\overline{H} = \overline{H(D)}$, which yields 1.24(c). □

Proof of Proposition 1.28. For the purpose of the proof, we denote by $\|\cdot\|_{\infty;[0,T]}$ either the supremum or the λ-essential supremum over $[0, T]$, depending on which of the two conditions in the statement of Proposition 1.28 holds.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The affine functions $\mathbb{R} \ni x \mapsto \alpha x + \eta$ with $\alpha, \eta \in \mathbb{R}$ are continuous over $[0,T]$ and therefore bounded. Since ψ is (locally λ-essentially) bounded, each $f \in \mathcal{NN}^d_{1,\infty}(\psi)$ is (λ-essentially) bounded, hence

$$\int_0^T f_i^2(s)\mu_{i,i}(\mathrm{d}s) \le \|f_i\|^2_{\infty;[0,T]} \int_0^T \pi_{i,i}(s)\,\mu(\mathrm{d}s) < \infty \tag{1.15}$$

for each $i \in \{1,2,\dots,d\}$. Note that in (1.15), we implicitly used the fact that μ is absolutely continuous with respect to λ in the case of Condition 1.27(3), since in this case each $f \in \mathcal{NN}^d_{1,\infty}(\psi)$ is μ-essentially bounded. We conclude that $\mathcal{NN}^d_{1,\infty}(\psi)$ is a linear subspace of $\Lambda^{2,0}$.

For $f \in \Lambda^{2,0}$, let $\epsilon > 0$. For each $\eta \in \mathbb{R}^d$ and $s \in [0,T]$, we have the inequality $\eta^\mathsf{T}\pi(s)\eta \le |\eta|^2 \operatorname{tr}(\pi(s))$ (cf. Cherny and Shiryaev (2002, Section 3)). By Lemma 1.10(d), the continuous functions $C([0,T];\mathbb{R}^d)$ are dense in $\Lambda^{2,0}$, hence there exists some $f_\epsilon \in C([0,T];\mathbb{R}^d)$ such that $\|f - f_\epsilon\|_{\Lambda^2} < \epsilon/2$. By Theorem 1.27, there exists some $g \in \mathcal{NN}^d_{1,\infty}(\psi)$ such that $\|f_\epsilon - g\|_{\infty;[0,T]} < \epsilon/(2\sqrt{\|\operatorname{tr}(\pi)\|_{L^1(\mu)}})$, hence

$$\|f - g\|_{\Lambda^2} \le \|f - f_\epsilon\|_{\Lambda^2} + \|f_\epsilon - g\|_{\Lambda^2} < \epsilon/2 + \|f_\epsilon - g\|_{\infty;[0,T]}\sqrt{\|\operatorname{tr}(\pi)\|_{L^1(\mu)}} < \epsilon,$$

which concludes our proof. □

Proof of Proposition 1.29. Since ψ is bounded, one can show precisely as in the proof of Proposition 1.28 that $\mathcal{NN}^d_{1,\infty}(\psi)$ is a linear subspace of $\Lambda^{2,0}$. If $\mathcal{NN}^d_{1,\infty}(\psi)$ were not dense in $\Lambda^{2,0}$, then there would exist by the geometric version of the Hahn–Banach theorem a functional $F \in (\Lambda^{2,0})^*$ such that $F \neq 0$ and $F(f) = 0$ for each $f \in \mathcal{NN}^d_{1,\infty}(\psi)$. Let N denote the subspace of all $G \in (\Lambda^2)^*$ that annihilate $\Lambda^{2,0}$, i.e. for which $G(f) = 0$ for each $f \in \Lambda^{2,0}$ holds. $(\Lambda^{2,0})^*$ can then be identified with the quotient space $(\Lambda^2)^*/N$.

From Lemma 1.10(b), we know that there exists a function $g \in \Lambda^2$, such that $F(f) = \int_0^T f^\mathsf{T}(s)\pi(s)g(s)\,\mu(\mathrm{d}s)$ for each $f \in \Lambda^{2,0}$. By linearity of the Lebesgue–Stieltjes integral,

$$F(f) = \sum_{i=1}^d \int_0^T f_i(s) \sum_{j=1}^d \pi_{i,j}(s)g_j(s)\,\mu(\mathrm{d}s) = 0, \quad f \in \mathcal{NN}^d_{1,\infty}(\psi), \tag{1.16}$$

and by a variant of the Cauchy–Schwarz inequality (cf. Cherny and Shiryaev (2002, Lemma 4.17)), for each $i \in \{1,2,\dots,d\}$ and $A \in \mathcal{B}_{[0,T]}$, it holds that

$$\int_A \left| \sum_{j=1}^d \pi_{i,j}(s)g_j(s) \right| \mu(\mathrm{d}s) \le \sqrt{\mu_{i,i}(A)}\|g\|_{\Lambda^2} \le \sqrt{\mu(A)}\|g\|_{\Lambda^2}, \tag{1.17}$$

hence $\nu_i(A) \coloneqq \int_A \sum_{j=1}^d \pi_{i,j}(s)g_j(s)\,\mu(\mathrm{d}s)$ defines a signed Borel measure on $[0,T]$ that is of finite total variation.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Exactly as in Cybenko (1989); Hornik (1991), we arrive at the question whether there exists a signed Borel measure $\nu \neq 0$ on $[0,T]$ that is of finite total variation, such that $\int_0^T \psi(\alpha x + \eta)\nu(\mathrm{d}x) = 0$ holds for all $\alpha, \eta \in \mathbb{R}$. As we know from Cybenko (1989, Lemma 1), this is not the case if ψ is bounded, measurable and sigmoidal (meaning that $\psi(t) \to 0$ as $t \to -\infty$ and $\psi(t) \to 1$ as $t \to \infty$), and Hornik (1991, Theorem 5) then generalized this finding to show that this is not the case if ψ is bounded, measurable and nonconstant. In other words, (1.16) implies that $F \equiv 0$, which yields a contradiction. □

Proof of Proposition 1.36. Since each $f \in D$ is a linear combination of compositions of ψ and affine functions, both of which are continuously differentiable, if follows that f is continuously differentiable, hence of bounded variation, which shows that $H(D)$ is a subspace of H_{bv}.

By Proposition 1.28, there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $H(D)$ that converges to $g_{h,M}$ in H. From the proof of Theorem 1.24(c) we know that the canonical injection from H to $C_0([0,T];\mathbb{R})$ is continuous, which implies that $(h_n)_{n\in\mathbb{N}}$ converges to $g_{h,M}$ in $C_0([0,T];\mathbb{R})$. Since $\tilde{F}\colon C_0([0,T];\mathbb{R}) \to \mathbb{R} \cup \{-\infty\}$ is assumed to be continuous, it follows that $\tilde{F}(h_n)$ converges to $\tilde{F}(g_{h,M})$. Moreover, since $\|\cdot\|_H\colon H \to \mathbb{R}_+$ is Lipschitz-continuous, we can conclude that $\tilde{F}_{h,M}(h_n)$ converges to $\tilde{F}_{h,M}(g_{h,M})$. □

The following lemma follows from standard arguments.

Lemma 1.51. *For all $h \in H$ and $p \in [1,\infty)$, we have $f_h \in L^p(M)$, which implies that $f_h^{\mathsf{T}} \bullet M \in \mathcal{H}^p$. Moreover, if $(h_n)_{n\in\mathbb{N}}$ denotes a sequence that converges to h in H, then $f_{h_n}^{\mathsf{T}} \bullet M \to f_h^{\mathsf{T}} \bullet M$ in $\mathcal{H}^p$ for each $p \in [1,\infty)$.*

Proof. Since f_h is deterministic it is, being interpreted as a stochastic process, predictable (cf. Schmock (2024, Exercise 7.103)). Moreover, since

$$\|f_h\|_{L^p(M)} = \mathbb{E}\big[[f_h^{\mathsf{T}} \bullet M]_T^{p/2}\big]^{1/p} = \mathbb{E}\big[((f_h^{\mathsf{T}} \pi f_h) \bullet C)_T^{p/2}\big]^{1/p} = \|h\|_H < \infty, \tag{1.18}$$

we have that $f_h \in L^p(M)$, and an application of the Burkholder–Davis–Gundy (BDG) inequality implies that $f_h^{\mathsf{T}} \bullet M \in \mathcal{H}^p$. Keeping in mind (1.18), an application of the BDG inequality then yields the existence of a positive constant c_p such that

$$\|(f_{h_n} - f_h)^{\mathsf{T}} \bullet M\|_{\mathcal{H}^p} \leq c_p \|h_n - h\|_H,$$

where the right-hand side converges to zero, as $n \to \infty$. □

Before we prove Lemma 1.37 let us recall for convenience a technical result, which follows e.g. from Elstrodt (2018, Satz 5.4) or by combining the proofs of Kallenberg (2021, Lemma 1.34) and Cohen and Elliott (2015, Theorem 1.3.39).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Lemma 1.52. *Fix $p > 0$, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $L^p(\mathbb{P})$ such that $f_n \to f$ in probability, where $f \in L^p(\mathbb{P})$. Then*

$$\|f_n - f\|_{L^p(\mathbb{P})} \to 0 \quad \Leftrightarrow \quad \|f_n\|_{L^p(\mathbb{P})} \to \|f\|_{L^p(\mathbb{P})}.$$

Proof of Lemma 1.37. Since both H and $L^p(\mathbb{P})$ are metric spaces, it suffices to prove the sequential continuity of A_p. Given $h \in H$, let $(h_n)_{n\in\mathbb{N}}$ be a sequence that converges to h in H, and set $Y_n = (A_p(h_n))^p$ for $n \in \mathbb{N}$. We will now show that $(Y_n)_{n\in\mathbb{N}}$ converges to $Y = (A_p(h))^p$ in $L^1(\mathbb{P})$. To this end, we apply Vitali's convergence theorem. Let us first collect some important properties.

(a) (convergence in probability) By Lemma 1.51, we have $f_{h_n}^{\mathsf{T}} \bullet M \to f_h^{\mathsf{T}} \bullet M$ in $\mathcal{H}^2$, hence $(f_{h_n}^{\mathsf{T}} \bullet M)_T \to (f_h^{\mathsf{T}} \bullet M)_T$ in $L^2(\mathbb{P})$ and thus also in probability. By the reverse triangle inequality, we have $\|h_n\|_H \to \|h\|_H$ as $n \to \infty$. An application of the continuous mapping theorem implies that $\exp\left(-p(f_{h_n}^{\mathsf{T}} \bullet M)_T + p\|h_n\|_H^2/2\right) = Y_n$ converges to $\exp\left(-p(f_h^{\mathsf{T}} \bullet M)_T + p\|h\|_H^2/2\right) = Y$ in probability.

(b) (boundedness in $L^1(\mathbb{P})$) For each $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}[Y_n] &= \mathbb{E}\big[\exp\big(-p(f_{h_n}^{\mathsf{T}} \bullet M)_T - p^2\|h_n\|_H^2/2\big)\big] \exp\left((p+p^2)\|h_n\|_H^2/2\right) \\ &= \mathbb{E}[Z_T^n] \exp\left((p+p^2)\|h_n\|_H^2/2\right),\end{aligned} \tag{1.19}$$

where $Z^n := \mathcal{E}(-p(f_{h_n}^{\mathsf{T}} \bullet M))$ is, by Novikov's criterion, a martingale. As a consequence, we have $\mathbb{E}[Z_T^n] = \mathbb{E}[Z_0^n] = 1$, hence the sequence $(Y_n)_{n\in\mathbb{N}}$ is bounded in $L^1(\mathbb{P})$, since we have already established that $\|h_n\|_H$ converges to $\|h\|_H$, which implies that $\exp\left((p+p^2)\|h_n\|_H^2/2\right)$ converges to $\exp\left((p+p^2)\|h\|_H^2/2\right)$.

(c) (uniform integrability) Fix $\varepsilon > 0$ and note that, by the same arguments that we used for Part (b),

$$\sup_{n\in\mathbb{N}} \mathbb{E}[(Y_n)^{1+\varepsilon}] < \infty.$$

De la Vallée Poussin's criterion implies that the set $\{Y_n : n \in \mathbb{N}\} \subset L^1(\mathbb{P})$ is uniformly integrable.

By Vitali's convergence theorem, we now have that $Y_n \to Y$ in $L^1(\mathbb{P})$. If we repeat the arguments laid out in Part (a) for $p = 1$, we see that $A_p(h_n) \to A_p(h)$ in probability. Moreover,

$$\|A_p(h_n)\|_{L^p(\mathbb{P})}^p = \|Y_n\|_{L^1(\mathbb{P})} \to \|Y\|_{L^1(\mathbb{P})} = \|A_p(h)\|_{L^p(\mathbb{P})}^p,$$

hence, Lemma 1.52 implies that $A_p(h_n) \to A_p(h)$ in $L^p(\mathbb{P})$. Finally, Eq. (1.19) shows

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

that $\|A_p(h)\|_{L^p(\mathbb{P})} = \exp((1+p)\|h\|_H^2/2)$, hence

$$\limsup_{\|h\|\to\infty} \frac{\|A_p\|_{L^p(\mathbb{P})}}{\|h\|_H} = \lim_{\|h\|\to\infty} \frac{\exp\big((1+p)\|h\|_H^2/2\big)}{\|h\|_H} = \infty,$$

which yields the remaining assertion. □

Proof of Theorem 1.38. As in the proof of Lemaire and Pagès (2010, Proposition 4), Hölder's inequality yields:

$$\begin{aligned} V(h) &= \mathbb{E}_{\mathbb{P}}\big[F^2(X)\exp\big(-(f_h^{\mathsf{T}} \bullet M)_T + \|h\|_H^2/2\big)\big] \\ &\le \mathbb{E}_{\mathbb{P}}\big[|F(X)|^{2+\varepsilon}\big]^{1/p}\mathbb{E}_{\mathbb{P}}\big[\exp\big(-q(f_h^{\mathsf{T}} \bullet M)_T + q\|h\|_H^2/2\big)\big]^{1/q} \\ &= \mathbb{E}_{\mathbb{P}}\big[|F(X)|^{2+\varepsilon}\big]^{1/p}\mathbb{E}_{\mathbb{P}}\big[\mathcal{E}\big(-q(f_h^{\mathsf{T}} \bullet M)\big)_T\big]^{1/q}\exp\big((1+q)\|h\|_H^2/2\big) \\ &= \mathbb{E}_{\mathbb{P}}\big[|F(X)|^{2+\varepsilon}\big]^{1/p}\exp\big((1+q)\|h\|_H^2/2\big), \quad h \in H, \end{aligned}$$

for $p = (2+\varepsilon)/2$ and $q = (2+\varepsilon)/\varepsilon$, where the last equality follows from the fact that $\mathcal{E}(-q(f_h^{\mathsf{T}} \bullet M))$ is actually a martingale, which implies that $\mathbb{E}_{\mathbb{P}}[\mathcal{E}(-q(f_h^{\mathsf{T}} \bullet M))_T] = \mathbb{E}_{\mathbb{P}}[\mathcal{E}(-q(f_h^{\mathsf{T}} \bullet M))_0] = 1$. This shows that V is $\mathbb{R}_+$-valued.

Next, let us show that V is convex. To this end, pick $\eta \in (0,1)$ and $g, h \in H$ such that $g \neq h$. By the triangle inequality and positive homogeneity, we have the inequality $\|\eta g + (1-\eta)h\|_H \le \eta\|g\|_H + (1-\eta)\|h\|_H$. By the linearity of the vector stochastic integral and the convexity of $\mathbb{R} \ni x \mapsto x^2$, we thus have

$$\begin{aligned} &-\big((\eta f_g + (1-\eta)f_h)^{\mathsf{T}} \bullet M\big)_T + \|\eta g + (1-\eta)h\|_H^2/2 \\ &\qquad \le \eta\big(-(f_g^{\mathsf{T}} \bullet M)_T + \|g\|_H^2/2\big) + (1-\eta)\big(-(f_h^{\mathsf{T}} \bullet M)_T + \|h\|_H^2/2\big). \end{aligned}$$

Together with the convexity and monotonicity of $\mathbb{R} \ni x \mapsto \exp x$, this shows that V is convex.

Given $h \in H$, let $(h_n)_{n\in\mathbb{N}}$ be a sequence that converges to h in H. Due to Lemma 1.37, $Z^n := A_q(h_n)$ converges to $Z := A_q(h)$ in $L^q(\mathbb{P})$. Note that $F^2(X) \in L^p(\mathbb{P})$ by assumption. By Riesz's representation theorem, the topological dual of $L^q(\mathbb{P})$ is isometrically isomorphic to $L^p(\mathbb{P})$, where the isomorphism is given by

$$L^p(\mathbb{P}) \ni g \mapsto \Big(L^q(\mathbb{P}) \ni f \mapsto \int_\Omega g(\omega)f(\omega)\mathbb{P}(\mathrm{d}\omega)\Big),$$

hence the map $L^q(\mathbb{P}) \ni Y \mapsto \mathbb{E}[F^2(X)Y] \in \mathbb{R}$ is continuous, which yields

$$\lim_{n\to\infty} V(h_n) = \lim_{n\to\infty} \mathbb{E}[F^2(X)Z^n] = \mathbb{E}[F^2(X)Z] = V(h).$$

Since H is in particular a metric space, continuity of V is equivalent to sequential

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

continuity, which shows 1.38(b).

In order to prove existence of a minimizer of V, we borrow some tools from convex optimization, see Zălinescu (2002) for a textbook treatment about this topic. First, we show that V is proper, meaning that $\{h \in H : V(h) < \infty\} \neq \varnothing$ and $V(h) > -\infty$ for all $h \in H$. The latter condition is clearly satisfied, as V is nonnegative. For $h \equiv 0$, we further have $V(h) = \mathbb{E}[F^2(X)] < \infty$, which implies the former condition (which also follows from 1.38(a)). Moreover, since V is continuous as argued above, it is in particular lower semicontinuous.

Let us show that V is coercive, i.e. that $V(h) \to \infty$, as $\|h\|_H \to \infty$. Since we assume that $\mathbb{P}[F^2(X) > 0] > 0$, there exists a constant $\delta > 0$ such that $\mathbb{P}[F^2(X) \geq \delta] > 0$. An application of the reverse Hölder inequality along the lines of the proof of Lemaire and Pagès (2010, Proposition 4) reveals the inequality

$$V(h) \geq \delta\, \mathbb{P}[F^2(X) \geq \delta]^3 \exp\left(\|h\|_H^2/4\right), \quad h \in H,$$

which shows that V is coercive.

Consequently, Zălinescu (2002, Proposition 2.5.6) shows that $\arg\min_{h\in H} V(h)$ is a convex set. Moreover, upon noting that H, being a Hilbert space, is reflexive, Zălinescu (2002, Theorem 2.5.1) shows that $\arg\min_{h\in H} V(h)$ is not empty, which shows 1.38(c).

For $\delta > 0$ choose $h_\delta \in H$ such that $V(h_\delta) < \min_{h\in H} V(h) + \delta$. By Theorem 1.24(a), there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $H(D)$ that converges to h_δ in H. By 1.38(b) we then obtain $\lim_{n\to\infty} V(h_n) = V(h_\delta) < \min_{h\in H} V(h) + \delta$. A diagonalization argument yields 1.38(d). □

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Part II.

Deep Surrogate Models

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

2. Life Cycle Insurance, Bequest Motives and Annuity Loads

In this chapter we investigate numerically the optimal purchases of insurance products in life cycle settings when bequest motives vary with age and life insurance and life annuities both carry loads. Our main research question is the "deep" annuity puzzle whereby a thin market for life annuities co-exists with a thick market for life insurance. The regular version of this puzzle restricts attention to the fact of a thin annuities market.

The life cycle model integrates an individual's life insurance phase with her life annuity phase, recognizing that buying a life annuity is like selling life insurance. This symmetry is evident when all life-contingent insurance is of the instantaneous-term variety, as is generally the case in the literature under discussion. The annuity buyer's estate makes a payment to the seller if and when the buyer dies. Likewise, the seller of life insurance makes a payment to the buyer's estate, contingent on the buyer's death. Yaari (1965), Hakansson (1969) and Merton (1971) led the way. Bequests were assumed to be normal goods without being either necessities or luxuries. In particular, constant relative risk aversion (i.e. power utility) was assumed whenever bequest utility had a specific functional form. Insurance was initially assumed to be actuarially fair.

Many subsequent contributions do incorporate loads. They generally specify loads in terms of markups over actuarially fair insurance. For this purpose, life cycle models typically introduce a single load parameter that can be interpreted as the ask price of life insurance. The parameter does double duty, for the life annuity phase as well as the life insurance phase. Early examples include the influential contributions of Fischer (1973) and Richard (1975). However, it turns out that the assumptions of a single-parameter model of loads and normal bequests both act to exaggerate the symmetry between the demand for life insurance and the demand for life annuities.

In the case of loads, we show below that the single-parameter load model of these early contributions (and many subsequent ones) entails negative annuity loads. As a consequence, the demand for annuities is artificially inflated. This is unhelpful for resolving the annuity puzzle, deep or otherwise. Moreover, it turns out that bid-ask spreads in the market for life-contingent insurance are effectively tied to zero. This implicit assumption leads to a counterfactual prediction that people will participate continuously in the market for life-contingent insurance.

Loads on life-contingent insurance products can be seen as a form of illiquidity or transaction cost. Such frictions have been studied in the context of risky securities. See

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

for example Magill and Constantinides (1976), Davis and Norman (1990) and Shreve and Soner (1994). These contributions analyze positive bid-ask spreads, in contrast to the existing literature on life-contingent insurance over the life cycle.

We introduce a second load parameter that can be interpreted as the bid price of life insurance. It ensures non-negative annuity loads and enables positive bid-ask spreads. Moreover, we assume bequests are necessities during child-rearing years. This is a natural assumption for people with dependents, notably, children at home. By contrast, bequests are luxuries during retirement years, consistent with a substantial recent body of research. Our main contribution, then, is to show numerically that a two-parameter load model and a specification of bequest motives that allows for age variation can jointly resolve the deep annuity puzzle. Previous contributions have not sought to account for the disparity between the demand for life insurance and the demand for life annuities.

Much of the life cycle literature is primarily concerned with life insurance. However, the dynamic programming principle suggests that parameters determining the demand for life annuities could feed forward onto the demand for life insurance. For completeness, we investigate this possibility. As it turns out, our computations suggest that changes in annuity loads have no effect on the demand for life insurance. Likewise, changes in life insurance loads have no effect on the demand for annuities.

The remainder of this chapter is organized as follows. Section 2.1 sets out an optimization problem. Section 2.2 compares and contrasts our approach with previous contributions to the 21st century literature. Section 2.3 presents a methodology that obtains reasonable estimates for loads. Subsection 2.4.1 reports results of our computations for the post-retirement phase and Subsection 2.4.2 reports results for the full life cycle. Subsection 2.5.1 reports results for the case of a truncated life cycle and Subsection 2.5.2 calibrates a Gompertz mortality model to data. Subsection 2.5.3 gives details of the age-earnings profile and the age-varying bequest parameter assumed here. Subsection 2.5.4 explains the dynamic programming approach used for the numerical solution of the optimization problem.

2.1. Model description

Consider a market for life-contingent insurance. The individual has the opportunity to purchase and sell term insurance on her own life, where the sale of insurance corresponds to the purchase of a life annuity. Life-contingent insurance is offered continuously, and the individual enters a contract by paying the premium rate $p(t)$, which buys life insurance in the amount $p(t)/\eta(t)$ or $p(t)/\theta(t)$ dollars for the next instant, depending on whether $p(t) > 0$ or $p(t) < 0$. The mortality rate, $\lambda(t)$, represents the instantaneous death rate for the individual surviving up to time t. Hence insurance is actuarially fair when $\theta(t) = \lambda(t) = \eta(t)$ for all t, and it is loaded when $\theta(t) \leq \lambda(t) \leq \eta(t)$, with at least one of those inequalities being strict for some time t. Under our two-parameter model of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

insurance loads, then, an individual with financial wealth $W(t)$ who dies at age t leaves a legacy $Z(t)$ given by

$$Z(t) = \begin{cases} W(t) + \frac{p(t)}{\eta(t)}, & \text{if } p(t) > 0, \\ W(t), & \text{if } p(t) = 0, \\ W(t) + \frac{p(t)}{\theta(t)}, & \text{if } p(t) < 0, \end{cases} \tag{2.1}$$

according as whether she is long, not invested or short in life insurance, the last case corresponding to being long in life annuities.

Bequest utility B is of hyperbolic absolute risk aversion (HARA) form with age-varying shift parameter $c_b(t)$,

$$B(t,z) = \left(\frac{\phi}{1-\phi}\right)^{\sigma} \frac{\left(\frac{\phi}{1-\phi} c_b(t) + z\right)^{1-\sigma}}{1-\sigma}, \tag{2.2}$$

where $\phi \in (0,1)$ is the marginal propensity to bequeath, and $\sigma > 0$ is the coefficient of relative risk aversion for consumption utility. The parameter $c_b(t)$ varies with age as described in Figure 2.1 below, where $c_b(t)$ is described as the age-dependency profile and is shown alongside $y(t)$, the age-earnings profile (or income function), which is standard; see e.g. Mincer (1974). Following Lockwood (2012, 2018) and others, the age-earnings profile incorporates an annuity component, reflecting a Social Security entitlement. Equations specifying the age-dependency profile and age-earnings profile are set out in Subsection 2.5.3.

Recent research finds that power utility is not a good model of the bequest utility of elderly individuals, especially affluent ones. Rather, bequests by the elderly tend to be luxury goods (Carroll, 1998; De Nardi et al., 2010; Lockwood, 2012, 2018). These contributions model the bequest motives of the elderly by means of a fixed positive value of $c_b(t)$ in Eq. (2.2). We extend their parametric model of bequest motives, into the individual's work years, which typically overlap substantially with child-rearing years. The additive shift parameter has a negative sign during that time. In more detail, bequests are necessities rather than luxuries during most of the individual's working life, becoming more essential in the event the family grows further, but eventually transiting to luxuries as dependents progressively attain independence. For example, Figure 2.1 postulates that $c_b(t)$ rises smoothly in value from about $-\$7,600$ at around age 40 (the assumed age of maximum dependency within the household) to \$32,900 at age 65. Following Lockwood (2018), we assume $c_b(t)$ is constant in time thereafter.

Since human capital has no closed-form solution here, necessities and luxuries must be characterized by the sign of the additive shift parameter in the bequest utility function rather than the proportionate response of planned bequests to a given increase in the sum of financial wealth and human capital. This alternative characterization is standard in

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

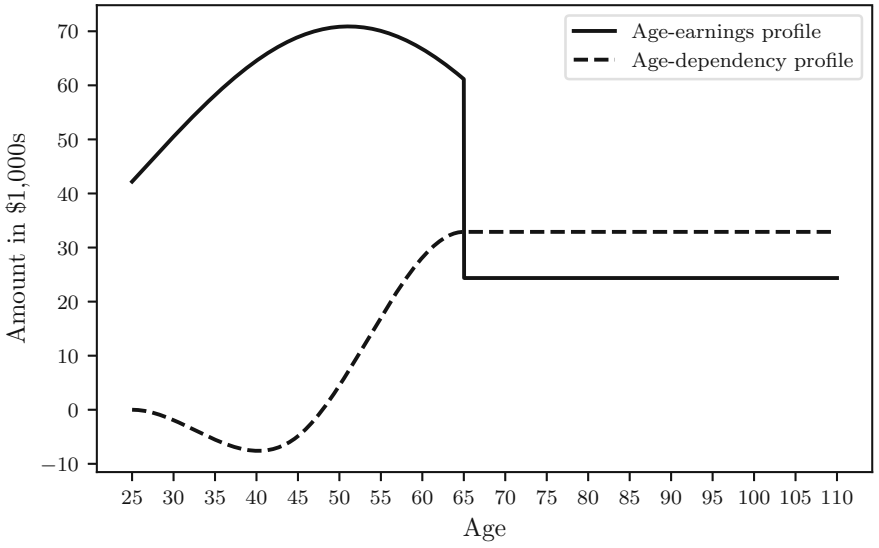

Figure 2.1. *Age-earnings profile and age-dependency profile.*

the literature on luxury bequests by the elderly. It readily extends to the pre-retirement phase. For comparison, we also report numeric results for the traditional life-cycle case of power bequest utility.

Assume that the individual's remaining lifetime, τ, is a random variable with known probability density function f. We introduce a finite planning horizon $T > 0$. The individual's maximum remaining lifespan need not coincide with T. For example, T could denote the time to retirement, as in Pliska and Ye (2007). The survival function $\bar{F}(t)$ modelling the probability of survival up to time t is $\bar{F}(t) = \exp(-\int_0^t \lambda(s)\,\mathrm{d}s)$. We assume a Gompertz model for $\lambda(t)$. For a functional form of $\lambda(t)$, see Table 2.1. The model was calibrated to mortality rates from the G12 countries; details are available in Subsection 2.5.2.

The individual consumes at rate $c(t)$. Instantaneous utility from consumption is of constant relative risk aversion (CRRA) form,

$$U(c) = \frac{c^{1-\sigma} - 1}{1 - \sigma}.$$

Given a pair (c, p) of consumption and insurance plans, the individual's financial wealth

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

$W(t)$ evolves according to

$$W(t) = w + \int_0^t \big(rW(s) + y(s) - c(s) - p(s)\big)\,\mathrm{d}s, \quad t \in [0, \tau \wedge T], \tag{2.3}$$

where w denotes the initial level of financial wealth, r denotes the risk-free rate, and $\tau \wedge T = \min\{\tau, T\}$. In case of death at time $t \leq T$, the total legacy $Z(t)$ is given by Eq. (2.1). If the individual survives beyond the planning horizon T, the legacy equals the financial wealth $W(t)$. As in previous contributions, we assume that the individual cannot leave a negative legacy. Given an initial level of financial wealth w, the individual's problem is to choose a consumption and insurance plan (c^*, p^*) that maximizes the expected utility of discounted consumption and bequest,

$$\begin{aligned}\sup_{(c,p)} \mathbb{E}\Big[\int_0^{\tau\wedge T} \mathrm{e}^{-\beta t} U\big(c(t)\big)\,\mathrm{d}t \\ + \mathrm{e}^{-\beta\tau} B\big(\tau, Z(\tau)\big) 1_{\{\tau \leq T\}} + \mathrm{e}^{-\beta T} B\big(T, W(T)\big) 1_{\{\tau > T\}}\Big],\end{aligned} \tag{2.4}$$

where β denotes the rate of time preference, $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to the randomness generated by the uncertain lifetime τ, and $1_{\{\cdot\}}$ denotes the indicator function. This is a problem that can be solved via dynamic programming; see Subsection 2.5.4. The model parameters are summarized in Table 2.1 below.

Table 2.1. *Model parameters.*

Model parameter	**Value**
Initial age of the individual	$x = 25$ years
Risk-free rate	$r = 3.2\%$
Risk-aversion coefficient	$\sigma = 2$
Rate of time preference	$\beta = -\ln(0.975)$
Propensity to bequeath	$\phi = 0.95$
Mortality rate	$\lambda(t) = \frac{1}{b}\exp\big(\frac{x+t-m}{b}\big)$
Modal age at death	$m = 88.23$
Scale parameter	$b = 9.38$

Note: Values for the risk-free rate, risk-aversion coefficient, rate of time preference and propensity to bequeath are from Lockwood (2018). Parameters for the mortality rate were calibrated to data from the G12 countries; see Subsection 2.5.2 below.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.2. Comparison with recent literature

Pliska and Ye (2007) make a leading theoretical and numeric contribution to this century's literature on life cycle insurance. Their theory offers a particularly clear exposition of that literature's single-parameter model of insurance loads. We therefore draw upon Pliska and Ye (2007) to explain how the ubiquitous single-parameter model gives rise to an implicit assumption of negative annuity loads and zero bid-ask spreads. They define $\eta(t) \geq \lambda(t)$ as the "premium-insurance ratio", that being the single load parameter in question. As in the preceding section, life-insurance premiums paid at the rate $p(t)$ secure for the upcoming instant a sum insured equal to $p(t)/\eta(t)$. Note that the parameter in the denominator can be interpreted as the ask price of life insurance. By assumption, life insurance carries a positive load whenever the inequality $\eta(t) \geq \lambda(t)$ is strict. So far so good. However, the single-parameter load model goes on to assume or imply that annuity income at the rate $p(t)$ entails an outlay of $p(t)/\eta(t)$ if the annuitant dies during the upcoming instant, payable by the annuitant's estate to the provider. If the inequality $\eta(t) \geq \lambda(t)$ is strict, this annuity is cheaper than an actuarially fair one. Annuities are implicitly subsidized. They make losses for insurance companies and will be in artificially high demand. For this reason, the preceding section introduced a second premium-insurance ratio, $\theta(t)$, $0 < \theta(t) \leq \lambda(t)$, for life annuities. If the individual dies at time t, her estate pays the provider an amount $p(t)/\theta(t)$. The parameter in the denominator can now be interpreted as the bid price of life insurance. Loads now remain non-negative when the individual goes long in life annuities. In this way, a two-parameter model of the (additive) bid-ask spread in the insurance market, namely $\eta(t) - \theta(t)$, can ensure that the spread is positive rather than zero.

Zero spreads in the context of a one-parameter model enable what is in effect a complete-markets solution, even when life insurance loads are positive. There is continuous participation in the market for life-contingent insurance. The reason is that although life insurance is more expensive than the actuarially-fair benchmark, this is just offset by life annuities being cheaper. Insurance demands, human capital amounts and the value function all have closed-form solutions. In these ways, relinquishing zero bid-ask spreads comes at the cost of reduced tractability. Once the bid-ask spread is positive, we lose effective market completeness and there is no longer continuous participation. We need to fall back on numerical computations.

Huang and Milevsky (2008) consider the case of HARA consumption utility whereby a state-varying shift parameter helps motivate the demand for life insurance. Bequests are effectively lifelong necessities. This life cycle setup is investigated both theoretically and numerically. The demand for annuities is found to be positive well before retirement, under either zero loads or loads described by a one-parameter model. People participate continuously in the market for life-contingent insurance.

Lockwood (2012, 2018) confines attention to the retirement phase. He addresses the annuity puzzle, combining luxury bequest utility with positive annuity loads. He

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

estimates the relevant shift parameter in the bequest utility function, and goes on to show numerically that the demand for annuity-type products is weak when there are luxury bequests in conjunction with realistic loads.

The calibrations reported by Pashchenko (2013) also deal exclusively with the retirement phase. She reports that only 5% of the singles aged 70 in her sample own a life annuity. By the same token, there is considerable dispersion in annuity purchases across income quintiles. An "administrative load" of 10% applies to annuity income. Loads also arise from adverse selection. There is a calibrated model of this source of loads on annuities, although the resulting total loads are not spelled out. Other factors depressing demand are luxury bequests, medical expenses, public annuity-like income, illiquid housing wealth, minimum purchase requirements and preannuitised wealth. Pashchenko's headline calibrated model predicts that 20% of singles aged 70 will participate in the market. In this way, Pashchenko (2013) accounts for most of her sample's non-participation in annuity markets. No single factor dominates the others as a prime cause of weak demand.

The classic overlapping generations model considers social welfare in an economy where bequests are not merely luxuries but provide no utility for the legator. Put another way, bequest motives are "inoperative". Even if the proceeds of unintended bequests are recycled by the government, the relative price of delayed consumption is too high, social saving is suboptimal, and capital formation is too low. Recent investigations of this scenario include Feigenbaum et al. (2013) and Heijdra et al. (2014). Our notion of necessity bequests during child-rearing years suggests that there may be offsetting tendencies at work. However, we leave this question to future research.

Peijnenburg et al. (2016) investigate the annuity puzzle in the context of a full (adult) life cycle. They show numerically that if (i) bequests are lifelong normal goods and (ii) insurance is actuarially fair, then the annuity puzzle persists under various circumstances that had previously been regarded as lessening it. By way of comparison, we show numerically that if (i) bequests evolve from being necessities to luxuries and (ii) all life-contingent insurance products carry a moderate positive load (e.g. 18% at age 65), then the annuity puzzle disappears, including the "deep" version of it.

2.3. Loads

Following Brown and Finkelstein (2007) and others, we specify loads as fixed markups over the benefits that would be payable were insurance actuarially fair. For example, Brown and Finkelstein (2007) find that a typical care insurance policy purchased by a 65 year old carries an 18% load. This means that for every dollar the individual pays in premiums, she expects care benefits worth 82 cents in present-value terms. Loads applying at age 65 together with Gompertz mortality serve to pin down lifelong load schedules. For this purpose we need to generalize the Brown–Finkelstein approach to incorporate life insurance. That turns out to be straightforward. For robustness, we consider a range

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

of loads for both life insurance and, especially, life annuities. We highlight the case of a single lifelong load schedule pinned down by an 18% load on both products when bought by individuals aged 65. This setup enables a parsimonious numeric resolution of the deep annuity puzzle.

Brown and Finkelstein (2007) confine attention to long-term care insurance policies covering home health care, assisted living and nursing home residence. That we treat their reported 18% load as our base case for term life annuities is in line with the estimate of Lockwood (2018). A load of 14% at age 65 is also noteworthy, as it turns out to be just sufficient to extinguish the demand for life annuities. A range of loads and their implications for the model parameters are shown in Figure 2.2 and Table 2.2 below.

Next, we present a methodology that allows us to obtain reasonable estimates for $\eta(t)$ and $\theta(t)$, starting from the assumption of an 18% load on insurance benefits for a 65 year old. For clarity, we henceforth use the subscripts *ins* and *ann*, depending on whether we are studying life insurance or life annuity loads.

2.3.1. Life annuity loads

Consider the expected net present value of a loaded, perpetual life annuity contract with continuous \$1 of payments, $\bar{a}(\kappa_{\text{ann}})$. Here, $\kappa_{\text{ann}} \geq 1$ denotes the mortality loading factor, with $1/\kappa_{\text{ann}}$ being applied as a multiplicative factor to the mortality rate $\lambda(t)$. The case $\kappa_{\text{ann}} = 1$ corresponds to no load, while $\kappa_{\text{ann}} > 1$ decreases mortality risk by reducing the mortality rate, thereby extending longevity and increasing the value of the life annuity contract.

Assuming a Gompertz model for $\lambda(t)$ with parameters from Table 2.1,

$$\begin{aligned}\bar{a}(\kappa_{\text{ann}}) &= \int_0^\infty \exp\Big(-rt - \int_0^t \tfrac{1}{\kappa_{\text{ann}}}\lambda(s)\,\mathrm{d}s\Big)\,\mathrm{d}t \\ &= \exp(C)\int_0^\infty \exp\Big(-rt - C\exp\big(\tfrac{t}{b}\big)\Big)\,\mathrm{d}t,\end{aligned}$$

where $C = \exp((x-m)/b)/\kappa_{\text{ann}}$. In order to bring the last integral into a more tractable form, we substitute $u = \exp(t/b)$,

$$\begin{aligned}\bar{a}(\kappa_{\text{ann}}) &= b\exp(C)\int_1^\infty u^{-(1+rb)}\exp(-Cu)\,\mathrm{d}u \\ &= b\exp(C)E_{1+rb}(C),\end{aligned}$$

where $E_s(z)$ denotes the generalized integro-exponential function (Milgram, 1985).

Assume that the individual purchases a life annuity contract by paying a single upfront premium P. For every \$1 of premium paid, the insurance company will only pay $\$(1 - L_{\text{ann}})$ of benefits, where $L_{\text{ann}} \in (0, 1)$ denotes a load. Given a load L_{ann}, we are

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

thus looking for a mortality loading factor κ_{ann}, such that

$$(1 - L_{\text{ann}})\,\bar{a}(\kappa_{\text{ann}}) = \bar{a}(1), \tag{2.5}$$

where $\bar{a}(1)$ is the expected net present value of benefits to be paid to the individual, and $P = \bar{a}(\kappa_{\text{ann}})$ is the premium that is charged by the insurance company.

Provided that Eq. (2.5) admits a solution, we define our premium-insurance ratio for the purchase of life annuities as $\theta(t) = \lambda(t)/\kappa_{\text{ann}}$ and observe that, by construction, the inequality $\theta(t) \leq \lambda(t)$ is always satisfied. Moreover, note that

$$\theta(t) = \frac{1}{b} \exp\Big(\frac{x + t - (m + \ln(\kappa_{\text{ann}}^b))}{b}\Big).$$

In other words, charging the premium-insurance ratio $\theta(t)$ corresponds to increasing the modal age at death from m to $m + \ln(\kappa_{\text{ann}}^b)$ in our model for the mortality rate $\lambda(t)$.

2.3.2. Life insurance loads

Consider the expected net present value of a loaded, perpetual life insurance contract with a payment of \$1 at the moment of death, $\bar{A}(\kappa_{\text{ins}})$. The mortality loading factor κ_{ins} is now being applied directly as a multiplicative factor to the mortality rate $\lambda(t)$. The case $\kappa_{\text{ins}} = 1$ corresponds to no load, while $\kappa_{\text{ins}} > 1$ increases mortality risk by increasing the mortality rate, thereby shortening longevity and increasing the value of the life insurance contract.

Assuming a Gompertz model for $\lambda(t)$ with parameters from Table 2.1,

$$\begin{aligned}\bar{A}(\kappa_{\text{ins}}) &= \int_0^\infty \exp\Big(-rt - \int_0^t \lambda(s)\kappa_{\text{ins}}\,\mathrm{d}s\Big)\lambda(t)\kappa_{\text{ins}}\,\mathrm{d}t \\ &= \frac{C}{b}\exp(C)\int_0^\infty \exp\Big(-(r - \tfrac{1}{b})\,t - C\exp(\tfrac{t}{b})\Big)\,\mathrm{d}t,\end{aligned}$$

where $C = \exp((x - m)/b)\kappa_{\text{ins}}$. In order to bring the last integral into a more tractable form, we again substitute $u = \exp(t/b)$,

$$\begin{aligned}\bar{A}(\kappa_{\text{ins}}) &= C\exp(C)\int_1^\infty u^{-rb}\exp(-Cu)\,\mathrm{d}u \\ &= C\exp(C)E_{rb}(C).\end{aligned}$$

Assume that the individual buys a life insurance contract by paying a single upfront premium P. Given a load L_{ins}, we are thus looking for a mortality loading factor κ_{ins}, such that

$$(1 - L_{\text{ins}}) \cdot \bar{A}(\kappa_{\text{ins}}) = \bar{A}(1), \tag{2.6}$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

where $\bar{A}(1)$ is the expected net present value of benefits to be paid to the individual, and $P = \bar{A}(\kappa_{\text{ins}})$ is the premium that is charged by the insurance company.

Provided that Eq. (2.6) admits a solution, we define our premium-insurance ratio for the purchase of life insurance as $\eta(t) = \lambda(t)\kappa_{\text{ins}}$ and observe that, by construction, the inequality $\lambda(t) \leq \eta(t)$ is always satisfied. Moreover, note that

$$\eta(t) = \frac{1}{b} \exp\Big(\frac{x + t - (m - \ln(\kappa_{\text{ins}}^b))}{b}\Big).$$

In other words, charging the premium-insurance ratio $\eta(t)$ corresponds to decreasing the modal age at death from m to $m - \ln(\kappa_{\text{ins}}^b)$ in our model for the mortality rate $\lambda(t)$.

2.3.3. Calibrating loads

We aim to find reasonable values of $\eta(t)$ and $\theta(t)$. Brown and Finkelstein (2007) estimate that the load for a policy purchased by a 65 year old is 18% if held until death. For this reason, we fix $x = 65$ as well as $L_{\text{ins}} = L_{\text{ann}} = 18\%$. Furthermore, we set 2% for the risk-free rate r used for discounting, a value that has also been used by Lockwood (2018) for calculating insurance premiums and benefits.

We solve Eq. (2.5) and Eq. (2.6) numerically via a root-finding algorithm, resulting in $\kappa_{\text{ann}} = 2.0377$ for life annuities, corresponding to an increase of the modal age at death to $m_{\text{ann}} = 94.91$, and $\kappa_{\text{ins}} = 4.7446$ for life insurance, corresponding to a decrease of the modal age at death to $m_{\text{ins}} = 73.63$.

Having fixed the mortality loading factors κ_{ann} and κ_{ins} to the values above, we compute the implied loads for ages other than $x = 65$ by solving Eq. (2.5) and Eq. (2.6) for L_{ins} and L_{ann} respectively. The evolution of those loads over time is presented in Figure 2.2. Fixed markups over fair insurance together with Gompertz mortality entail loads that vary with the buyer's age, as noted by Brown and Finkelstein (2007), among others. The curve for life annuities qualitatively resembles Figure 1 in Brown and Finkelstein (2007). While the differences between the two figures might be explained by differing assumptions about the risk-free rate r (we use $r = 2\%$ to match the setting of Lockwood (2018), while Brown and Finkelstein (2007) use the term structure of yields of U.S. Treasury strips), we conclude that our model for premium-insurance ratios is broadly consistent with practice.

The choice of load L has a direct impact on the premium-insurance ratios $\eta(t)$ and $\theta(t)$. Table 2.2 sets out mortality loading factors and their implied modal ages at death which were calibrated at the age 65 for different loads L. These are inputs to Table 2.3, which sets out annuity demands corresponding to different loads and ages.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

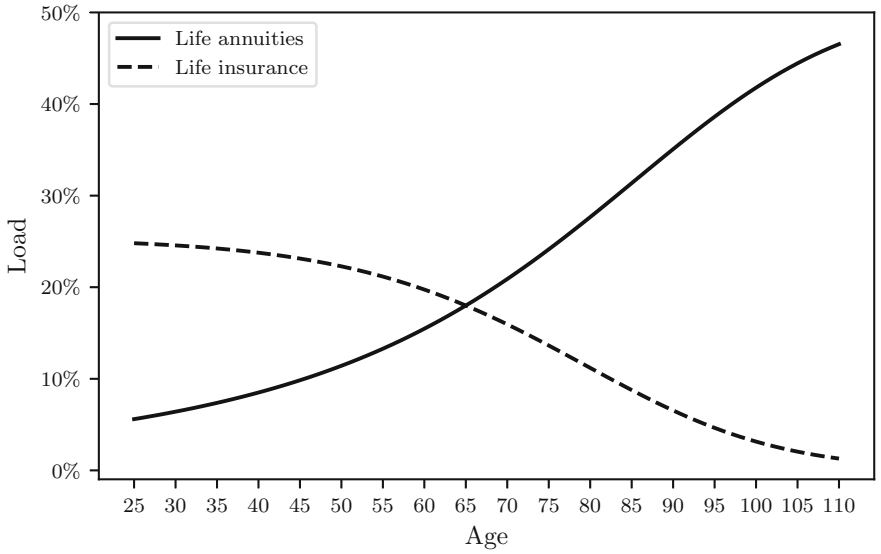

Figure 2.2. *Insurance loads L_{ann} and L_{ins} by age of purchase, with fixed mortality loads $\kappa_{\text{ann}} = 2.0377$ and $\kappa_{\text{ins}} = 4.7446$.*

Table 2.2. *Mortality loading factors κ_{ins}, κ_{ann} and corresponding implied modal ages at death m_{ins}, m_{ann} for different choices of load L.*

L	κ_{ins}	m_{ins}	κ_{ann}	m_{ann}
0%	1.0000	88.23	1.0000	88.23
2%	1.1482	86.93	1.0678	88.85
4%	1.3264	85.58	1.1434	89.49
6%	1.5426	84.16	1.2280	90.16
8%	1.8081	82.67	1.3232	90.86
10%	2.1381	81.10	1.4306	91.59
12%	2.5547	79.43	1.5527	92.36
14%	3.0903	77.65	1.6921	93.16
16%	3.7941	75.72	1.8523	94.01
18%	4.7446	73.63	2.0377	94.91
20%	6.0742	71.31	2.2537	95.85

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

2.4. Numerical study

2.4.1. After retirement

Consider financial plans made (or remade) at the age of retirement, assumed here to be 65. Fischer (1973) showed that CRRA bequest utility generates substantial demand for annuities, even before retirement. Annuity loads were (implicitly) negative in that study.

Figure 2.3 shows results of our computations in the case of an 18% load and financial wealth of $500,000 at the time of retirement. It suggests the following observations. CRRA bequest utility and fair annuities together generate a strong demand for annuities that begins before retirement (consistent with previous contributions) and lasts for as long as the maximum lifespan of the retiree. The demand is particularly strong at advanced ages. Luxury bequest utility and fair annuities together see a considerable fall in the demand for annuities, although demand still begins before retirement and lasts for as long as the maximum lifespan of the retiree. CRRA bequest utility and an 18% load on annuities together see not only a further fall in the demand for annuities, but non-participation in the market, beginning at age 97. Intuition for non-participation is given in the next section. Finally, luxury bequest utility and an 18% load see negligible participation in the market for annuities, in line with the findings of Lockwood (2012, 2018).

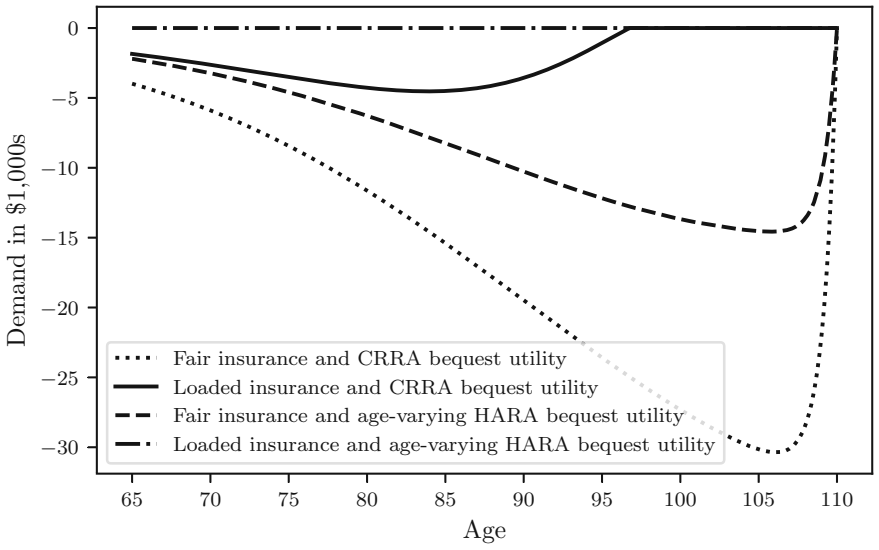

Figure 2.3. *Optimal life insurance and life annuity purchase after retirement.*

A more granular approach reveals the load value that just suffices to extinguish annuity

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

demand in the case of luxury bequests and positive loads. In particular, Table 2.3 shows that a 14% load just extinguishes annuity demand in this case. The table also shows that, as we progressively reduce loads below 14%, the demand for annuities at advanced ages rises to substantial levels, even with luxury bequests. Demand is sensitive to loads. As a robustness check, we also calibrated annuity demands when wealth at the time of retirement is either reduced to \$400,000 or increased to \$600,000. Broadly speaking, there are commensurate falls and rises in annuity demands. For example, when the load is 10% and wealth falls from \$500,000 to \$400,000, annuity demand by an individual aged 65 falls from \$480 per year to \$400 per year.

Table 2.3. *Load-dependent annuity demand (in* \$100*s) at different ages when financial wealth at the time of retirement is* \$500,000 *and bequests are luxuries.*

L	65	70	75	80	85	90
0%	22.1	32.4	46.0	62.9	82.4	102.8
2%	18.3	26.2	35.7	45.9	54.2	56.0
4%	14.7	20.2	25.9	29.8	28.0	13.8
6%	11.2	14.5	16.6	14.8	4.0	0.0
8%	7.9	9.1	7.9	0.9	0.0	0.0
10%	4.8	4.0	0.0	0.0	0.0	0.0
12%	1.9	0.0	0.0	0.0	0.0	0.0
14%	0.0	0.0	0.0	0.0	0.0	0.0

2.4.2. Full life cycle

Consider financial plans made at the outset of working life and family formation, assumed for simplicity to coincide at age 25. Fischer (1973) pioneered numeric studies of life cycle models of the demand for life insurance and life annuities when loads are of the one-parameter variety.

Bequest utility here is either CRRA (normal bequests) or HARA with an age-varying shift parameter (necessity bequests that transition to luxury bequests). Figure 2.4 portrays the case of an 18% load on both life insurance and life annuities, and compares it to the case of zero loads. Positive annuity loads and age-varying bequest jointly modify the demand for life-contingent insurance over the life cycle, to the point where decades-long spans of non-participation open up. In more detail, fair insurance ensures continuous participation regardless of whether bequest utility is CRRA or age-varying HARA.

A load of 18% combined with CRRA utility sees two periods of non-participation, one running from age 41 to 44, and the other beginning at age 97. Intuition for the midlife non-participation period can be gained from observing that fair insurance combined

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

with CRRA utility sees the crossover from life insurance to life annuities occur between ages 41 and 44. Once the load is introduced, non-participation seeps out from either side of the fair-insurance cutpoint. Similarly, for the late-life non-participation period, non-participation spreads to the left from the maximum-lifespan cut-point, namely, age 110.

Loaded insurance and age-varying bequests together see lifelong non-participation from age 51. It is evident that realistic loads can have strong effects, especially on annuity demands in conjunction with luxury bequests for retirees. Moreover, the demand for life insurance is weak when insurance is not a necessity during some working years. The figure also shows that age-varying bequest motives combine with loads to explain the deep annuity puzzle.

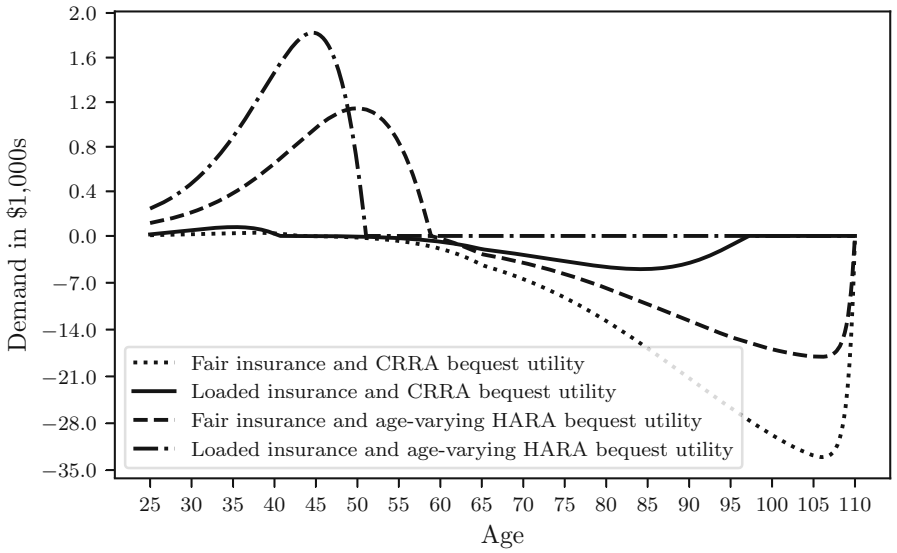

Figure 2.4. *Optimal life insurance and life annuity purchase over the full life cycle.*

Next, we examine the effects of annuity loads on the demand for life insurance, see Figure 2.5. In particular, we consider loads of either 6%, 12% or 18% while holding the life insurance load fixed at 12%. Bequest utility is age-varying HARA. Our computations suggest that there is no effect of changes in annuity loads on the demand for life insurance. For example, when the annuity load is raised from 12% to 18%, there is a fall in the demand for life annuities accompanied by a lengthening of the midlife non-participation period. However, the age at which non-participation begins remains unchanged at age 53. Likewise, there is no change in the life-insurance profile. Part of the intuition is that

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

life-contingent insurance here is of the instantaneous-term variety, so that individuals only look an instant ahead when making insurance decisions.

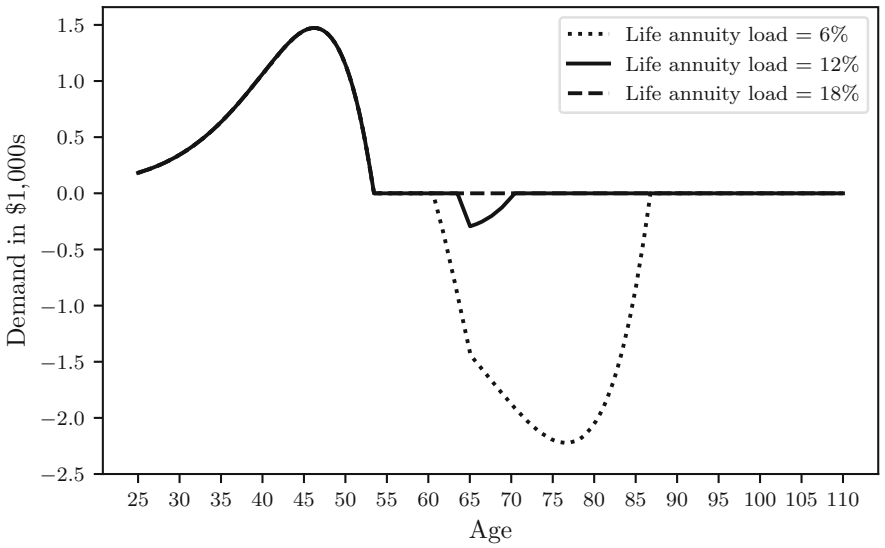

Figure 2.5. *Effects of different annuity loads on the demand for life insurance.*

Finally, consider the effects of changes in life insurance loads on the demand for life annuities. Bequest utility is again age-varying HARA. Paralleling the results portrayed in Figure 2.5, there is no feedback effect, regardless of life insurance loads, see Figure 2.6. Take the case of a 12% load. The corresponding annuitization phase begins at age 64 and ends at age 70. It is not affected by changes in life insurance loads. This figure also shows that the demand for life insurance is not highly sensitive to loads, in contrast to the demand for annuities.

Conclusions

Numerous previous investigations have examined the optimal demand for life insurance and life annuities in life cycle settings and under insurance loads. Loads have been modelled by means of a single parameter that effectively eliminates bid-ask spreads in insurance markets. This helps explain why numeric investigations have typically predicted continuous participation in insurance markets along with strong demand for annuities. A two-parameter model can restore positive bid-ask spreads. Positive spreads render markets incomplete, making it necessary to examine optimal insurance purchases by means of computations.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

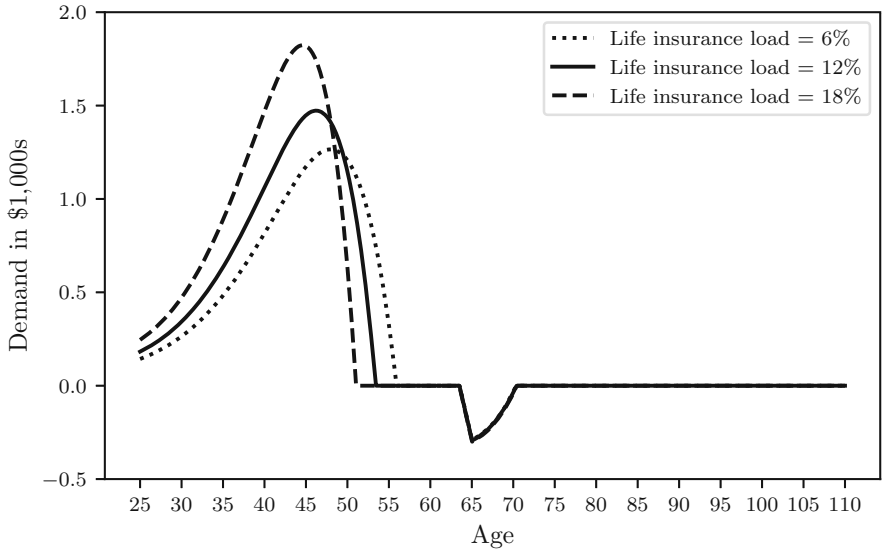

Figure 2.6. *Effects of different insurance loads on the demand for life annuities.*

Our computations suggest that positive annuity loads induce up to two periods of non-participation, one in midlife and the other adjoining the maximum age at death. Luxury bequests are not a necessary condition for two (rather than one) non-participation periods.

When bequests are luxuries, loads can have big effects on annuity demands. According to our computations for someone retiring at age 65, if the load is 14% or more of premiums then the demand for annuities is negligible. More generally, our computations suggest that the demand for loaded annuities is comparatively weak around age 65 even when bequests are not luxuries. By contrast, the demand for fair annuities is comparatively strong even when bequests are luxuries, especially in advanced old age.

To the extent bequests are necessities, loads no longer have large negative effects on insurance demands. Life insurance is a case in point. According to our computations for a 25 year old, for whom bequests are a necessity (while in gradual transit to being luxuries), the demand for life insurance (calibrated by reference to an 18% load at age 65) is strong. Moreover, compared to fair insurance, a life insurance load of 18% (calibrated to age 65) sees a higher peak demand for life insurance. This peak occurs at age 45. In other words, life insurance is a Giffen good in mid working life. This echoes a numeric finding due to Pliska and Ye (2007) – see their Figures 5 and 6. According to our Figure 2.6, an 18% load sees the demand for life insurance peter out at age 51. Moreover, the same load on annuities sees a negligible demand for them. A maximum lifespan of 110 years therefore

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

implies that the period of non-participation in the market for life-contingent insurance could be up to 59 years in length.

Our life-cycle computations suggest that changes in annuity loads do not affect life insurance decisions, although such changes do alter the age at which midlife non-participation ends. Likewise, changes in life insurance loads do not affect the demand for annuities.

Our computations need to be interpreted with caution. First, post-retirement parameters were based on Lockwood (2018), who derives his estimates from the Health and Retirement Study. As a consequence of Social Security, there is some pre-annuitization of retirement wealth. Second, adverse-selection issues make instantaneous-term products difficult for annuity providers, especially in the case of potential customers at advanced ages. Finally, there is ample scope for further investigation of bequest motives during working life. These will vary from one household to the next. For example, our assumed age-dependency profile attained a minimum at age 40, corresponding to a maximum desire to provide for dependents. Yet this aspect of preferences is highly idiosyncratic across households. Estimating the bequest-utility parameters of working-age people is a topic we leave to future research.

2.5. Supplementary material

2.5.1. Truncated life cycle

Some investigators focusing on life insurance collapse the retirement phase into a single point in time. Numeric studies with single-parameter models of loads and financial plans terminating with the individual's retirement (rather than her maximum lifespan) were pioneered by Pliska and Ye (2007). They found that (subsidized) annuity demands over this truncated horizon were modest, though discernible. Participation was found to be continuous. We redo their Figure 6, including here the case of positive (rather than negative) annuity loads.

In Figure 2.7, the planning horizon can be seen from the horizontal axis to be 40 years. Demands for life insurance and life annuities are shown on the vertical axis. The dotted line portrays the base case of Pliska and Ye's numeric analysis: loads are zero, the coefficient of relative risk aversion is 4, the rate of time preference is 3% per year, and the discount rate is 4% per year. The hazard rate is assumed to be a simple linear function of age – see Pliska and Ye (2007) for details. There is no annuity phase in the base case. The dashed line shows demands under a single-parameter model of the premium-insurance ratio. That case corresponds to $\kappa_{\text{ins}} = 4$ and $\kappa_{\text{ann}} = 1/4$. Introducing subsidized annuities (along with loaded life insurance) induces a positive annuity phase. It is about one year in length. Finally, the solid line shows demands under a two-parameter model, thereby ensuring that life insurance and life annuities both carry loads. The premium-insurance ratio corresponds to $\kappa_{\text{ins}} = 4$ during the insurance phase and $\kappa_{\text{ann}} = 4$ thereafter. In place of an annuity phase there is a non-participation period. It is about

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

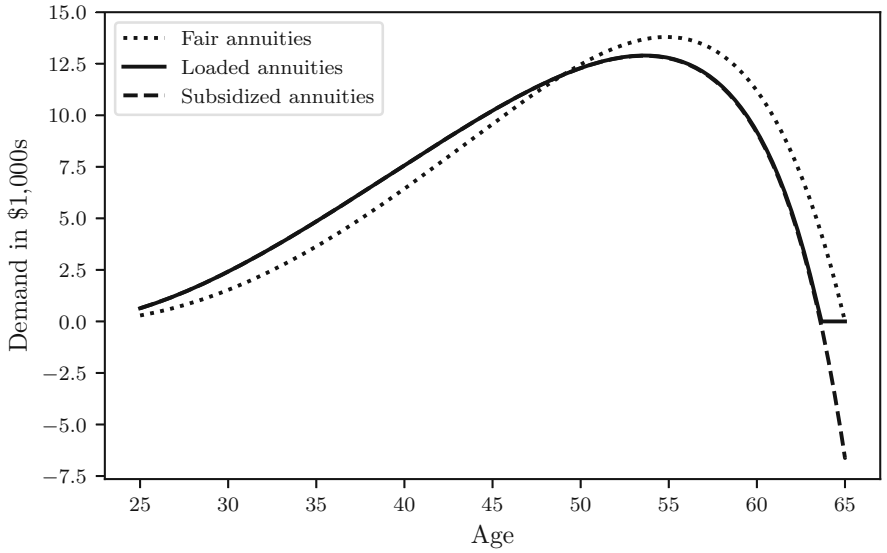

Figure 2.7. *Positive annuity loads in the Pliska–Ye model.*

one year in length and adjoins the retirement time $T = 40$. As with zero loads, there is no annuity phase in this case.

2.5.2. Estimating a Gompertz mortality model

We calibrate a model for the mortality rate $\lambda(t)$ to data. Our sample consists of life tables from the 2019 cohort of the G12 countries (Australia, Belgium, Canada, France, Germany, Italy, Japan, Netherlands, Spain, Sweden, Switzerland, UK and USA), which was obtained from the Human Mortality Database (2022). For each country $i = 1, 2, \ldots, 13$, and age $x = 25, 26, \ldots, 110$, we have data on l_x^i, the number of survivors at age x, and d_x^i, the number of deaths within the subsequent year from those survivors aged x years. The age 110 is not a single-year age, being treated as a cemetery state with $l_{110}^i = d_{110}^i$. We aggregate both the number of survivors and the number of deaths over all countries, and denote the corresponding age-dependent aggregates l_x and d_x, respectively.

We aim to estimate the parameters b and m of the mortality model set out in Table 2.1. We proceed as follows:

- Approximate the survival function via the Levenberg–Marquardt nonlinear least-squares method;
- Calibrate the mortality rate via the maximum likelihood method.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Both the survival function as well as the mortality rate enter directly into the definition of the value function V according to the dynamic programming principle in Subsection 2.5.4 below, so that it is important to have a good estimate for both functions.

Fitting the survival function

Under the parametric model for $\lambda(t)$ set out in Table 2.1, the survival function $\bar{F}(t)$ takes the form

$$\bar{F}(t) = \exp\Big(-\exp\big(\tfrac{25-m}{b}\big)\big(\exp\big(\tfrac{t}{b}\big) - 1\big)\Big).$$

It admits the representation $\bar{F}(t) = l_{25+t}/l_{25}$. We have information about the aggregated number of survivors l_{25+t} for $t = 0, 1, \ldots, 85$. We thus have 86 training points and aim to find parameters $\hat{b}$ and $\hat{m}$ such that the error

$$\frac{1}{2}\sum_{t=0}^{85}\Big(\bar{F}(t) - \frac{l_{25+t}}{l_{25}}\Big)^2$$

is minimized. This is done via the Levenberg–Marquardt nonlinear least squares method; see Table 2.4 for the estimated coefficients.

Maximum likelihood estimation

We follow Lenart (2014) and assume d_{25+t} follows a Poisson distribution,

$$d_{25+t} \sim \text{Poisson}(E_{25+t}\,\lambda(25+t)),$$

where E_{25+t} denotes the total number of person-years exposed to death at age $25+t$, defined analytically via $E_{25+t} = \int_0^1 l_{25+t+u}\,\mathrm{d}u$. By construction it follows that $l_{25+t} \geq E_{25+t} \geq l_{25+t+1}$ for all t. E_{25+t} accounts for the fact that not all l_{25+t} lives are exposed to mortality risk throughout the whole period $[25+t, 25+t+1)$, but only for a sub-period until death.

For simplicity, we assume that

$$E_{25+t} = l_{25+t} - (1 - a_{25+t})d_{25+t}, \quad t = 0, 1, ..., 84,$$

where a_{25+t} denotes the average number of years lived within the age-interval $[25+t, 25+t+1)$ for those people dying at that age, which we set to $a_{25+t} = 0.5$ for $t = 0, 1, ..., 84$. This is in line with the methodology used for the construction of life tables in the Human Mortality Database (2022).

As already mentioned, the age $x = 110$ is special in our sample, because it is not a single-year age, but a cemetery state, where $l_{110} = d_{110}$. In order to avoid introducing a special treatment and therefore a modelling bias, we exclude this state for the remainder

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

of this subsection, and estimate our parameters only based on the ages $x = 25, 26, ..., 109$. Note that using the cemetery age $x = 110$ for approximating the survival function in the previous subsection is not a problem though, because l_{110} still gives the proper interpretation of those lived at the beginning of the cemetery age.

We re-parameterize the mortality rate,

$$\lambda(t) = \tfrac{1}{b}\exp\left(\tfrac{25-m}{b}\right)\exp\left(\tfrac{t}{b}\right) = \alpha\exp(\beta t),$$

where $\alpha = \exp((25-m)/b)/b$ and $\beta = 1/b$. Following Lenart (2014), the maximum likelihood estimator $\hat{\alpha}$ for α is

$$\hat{\alpha} = \frac{\sum_{t=0}^{84} d_{25+t}}{\sum_{t=0}^{84} E_{25+t}\exp(\beta t)},$$

and the maximum likelihood estimator $\hat{\beta}$ for β is the root of the function

$$f(y) = \frac{\sum_{t=0}^{84} d_{25+t}t}{\sum_{t=0}^{84} d_{25+t}} - \frac{\sum_{t=0}^{84} E_{25+t}\exp(yt)t}{\sum_{t=0}^{84} E_{25+t}\exp(yt)}.$$

Conveniently, this representation allows for a straightforward calculation of the first derivative,

$$f'(y) = \left(\frac{\sum_{t=0}^{84} E_{25+t}\exp(yt)t}{\sum_{t=0}^{84} E_{25+t}\exp(yt)}\right)^2 - \frac{\sum_{t=0}^{84} E_{25+t}\exp(yt)t^2}{\sum_{t=0}^{84} E_{25+t}\exp(yt)},$$

and we can find the root of f numerically, as suggested by Lenart (2014), via the Newton–Raphson method; see Table 2.4 for the estimated coefficients.

Numerical calibration

We estimate the Gompertz mortality parameters b and m via both the Levenberg–Marquardt and the maximum likelihood method. For the final model coefficients, we choose a weighted average, where 25% of the weight is allocated to the first method (fitting the survival function), and 75% is allocated to the second method (maximum likelihood estimation). The reason for this weighting is that for our optimal control problem, both the survival function and the mortality rate appear in the definition of the value function. However, since the mortality rate $\lambda(t)$ plays a more prominent role through the definition of the premium-insurance ratios $\eta(t)$ and $\theta(t)$, we overweight the second set of estimated parameters.

Figure 2.8 portrays our estimated model, the aggregated mortality rate, and the mortality rates of each individual constituent country of the G12 group. The mortality rates are defined via $m_x^i = d_x^i/E_x^i$ and $m_x = d_x/E_x$, where E_x^i, the country-level total

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 2.4. *Estimated Gompertz mortality model parameters.*

Method	$\hat{b}$	$\hat{m}$
Levenberg–Marquardt	9.45	88.79
Maximum likelihood	9.35	88.05
Weighted average	9.38	88.23

Note: Reported values are rounded to two decimal places.

number of person-years exposed to death at age x were obtained similarly as above by replacing the aggregate values by country-level values.

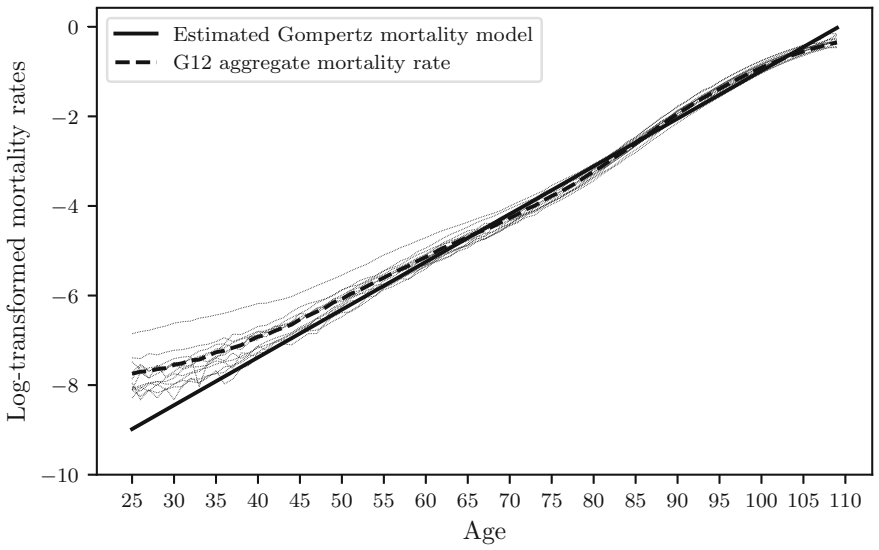

Figure 2.8. *Mortality rates of G12 countries, their aggregate and Gompertz model fit.*

2.5.3. Age-earnings and age-dependency profiles

We follow Mincer (1974), who discusses a quadratic form of the income function fit to the logarithm of weekly earnings. All dollar values in this section were rounded generously.

The average income of single retirees in Lockwood's (2018) sample is \$18,360, expressed in constant 2010 dollars using the Consumer Price Index for Urban Wage Earners and

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Clerical Workers (CPI-W). The annual average CPI-W stood at 213.967 points in 2010 and 283.926 points in 2022 (average over the first half of the year). This corresponds to an increase of approximately 32.69%. Adjusting Lockwood's income figure, we therefore assume a post-retirement income of \$24,360. At the moment of retirement, we postulate a drop in income of 60%, which implies that income immediately pre-retirement is \$60,900.

To generate more data points, we assume that income grows by 35% in the first 10 years in the labor market, by 25% in the next 15 years, and declines by 15% in the final 15 years (Murphy and Welch, 1990). We thus arrive at the following data points:

- The initial income is \$42,460;
- After 10 years in the labor market, the income is \$57,320;
- The peak occurs after 25 years with an income of \$71,650;
- Immediately before retirement, the income is \$60,900.

We fit an age-income profile $y(t)$ to these data points, leading to

$$y(t) = \exp(-0.000763\,t^2 + 0.0398\,t + 10.65).$$

Turning to the age-dependency profile, Lockwood estimates $c_b(t)$ post-retirement to be \$24,800. Adjusting Lockwood's value to constant 2022 dollars yields the value of \$32,900. Our post-retirement income is \$24,360. Therefore, $c_b(t)$ post-retirement is approximately 35% larger than post-retirement income. For symmetry reasons, we postulate that pre-retirement, minimum bequest after 20 years in the labor market should be approximately 35% larger than the income at that time. Note that, due to our choice of utility function $B(t, z)$ for the case $c_b(t) < 0$, the minimum bequest is $-\phi/(1-\phi)c_b(t)$, so we have to account for the scaling with $\bar{\phi} = \phi/(1-\phi)$. We thus arrive at the following data points: $c_b(0) = 0$, $c_b(20) = -1.35\,y(20)/\bar{\phi}$, and $c_b(40) = 32{,}900$.

We fit piecewise cubic splines over the intervals $[0, 20]$ and $[20, 40]$ with natural boundary conditions, meaning that the first derivatives are set to zero at the boundaries of those two intervals. This yields approximately the following specification for $c_b(t)$, where for ease of presentation we rounded the coefficients to two decimal places:

$$c_b(t) = \begin{cases} 4.31\,t^3 - 98.42\,t^2 & t \in [0, 20), \\ \quad - 6.36\,(t-20)^3 + 160.11\,(t-20)^2 & \\ \qquad\qquad + 1{,}233.75\,(t-20) - 4{,}897.43 & t \in [20, 40), \\ 32{,}900 & t \in [40, 85]. \end{cases}$$

2.5.4. Dynamic programming

For completeness, we set out the dynamic program, leading to a numerical algorithm that can be used to compute optimal solutions; see Algorithm 1. Here we treat a problem of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

deterministic optimal control. For a textbook treatment, see Fleming and Soner (2006).

In assuming a Gompertz mortality model for $\lambda(t)$ as specified in Table 2.1, the density f of the individual's remaining lifetime τ has the form

$$f(t) = \lambda(t)\bar{F}(t) = \frac{1}{b}\exp\Big(C + \ln(C) + \tfrac{t}{b} - C\exp(\tfrac{t}{b})\Big),$$

where $C = \exp((25-m)/b)$. From this we can see that $f(t) \to 0$ as $t \to \infty$, because the term $C\exp(t/b)$ dominates the term t/b inside the exponent. However, for each $t > 0$, we also have $f(t) > 0$ as well as $\bar{F}(t) > 0$. In other words, assuming a Gompertz mortality model for $\lambda(t)$ implies that for any time $t > 0$, there is a positive probability that the individual will survive up to this time.

Richard (1975) avoided this case by assuming that the support of the distribution of τ is bounded from above by a finite constant, say $\tilde{T} > 0$. Choosing $T = \tilde{T}$ would then yield a proper life cycle model. One disadvantage of this approach is that one cannot use the Gompertz law of mortality, which is popular in actuarial practice. Alternatively, one could truncate the distribution of τ to a finite interval $[0, \tilde{T}]$ by replacing the density $f(t)$ above with $\tilde{f}(t) = f(t)/F(\tilde{T})1_{[0,\tilde{T}]}(t)$, where $F(t)$ denotes the cumulative distribution function of τ. In that case, we would have

$$\mathbb{P}(\{\tau \le \tilde{T}\}) = \int_0^{\tilde{T}} \tilde{f}(t)\,\mathrm{d}t = \frac{F(\tilde{T})}{F(\tilde{T})} = 1.$$

It would be possible to replace the Gompertz mortality model by another model satisfying $\bar{F}(t) = 0$ for all t above a reasonable time $\tilde{T}$, say $\tilde{T} = 85$, or by truncating the model as described above to obtain a distribution over $[0, \tilde{T}]$. However, we note that for the chosen model parameters as specified in Table 2.1, the probability that a 25 year old individual survives past the year 110 is $\bar{F}(85) \approx 3.78 \cdot 10^{-5}$. Therefore, the Gompertz mortality model with parameters as described in Table 2.1 and with $T = 85$ is an adequate approximation to optimal consumption and premium plans over the full life cycle. This approximation omits only survival scenarios with a negligible probability of occurrence.

Given an initial level of wealth w, time $s < T$ as well as a pair (c, p) of consumption and insurance plans, let

$$J(s,w;c,p) = \mathbb{E}\Big[\int_s^{\tau\wedge T} \mathrm{e}^{-\beta(t-s)}U(c(t))\,\mathrm{d}t + \mathrm{e}^{-\beta(\tau-s)}B(\tau,Z(\tau))1_{\{\tau\le T\}} + \mathrm{e}^{-\beta(T-s)}B(T,W(T))1_{\{\tau>T\}}\,\big|\,\tau > s\Big],$$

where $\mathbb{E}[\,\cdot\,|\,\tau > s]$ denotes the expectation conditional on $\{\tau > s\}$. The only source of randomness is the uncertain lifetime τ. Following an approach pioneered by Yaari (1965),

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

$J(s, w; c, p)$ can be rewritten:

$$J(s,w;c,p) = \int_s^T \mathrm{e}^{-\beta(t-s)}\big(\bar{F}(t,s)U(c(t)) + f(t,s)B(t,Z(t))\big)\,\mathrm{d}t \\ + \mathrm{e}^{-\beta(T-s)}\bar{F}(T,s)B(T,W(T)),$$

where $\bar{F}(t,s) = \bar{F}(t)/\bar{F}(s)$ denotes the conditional survival probability, and $f(t,s) = f(t)/\bar{F}(s)$ denotes the conditional probability density.

We can now use deterministic dynamic programming over the fixed time horizon $[0,T]$. We assume that $y(t)$ is integrable (which is the case for the age-income profile specified in Subsection 2.5.3). Moreover, we restrict our attention to consumption and premium plans (c,p) which are bounded and measurable. Then Eq. (2.3) is well defined. To further assure admissibility, for each w and $s \in [0,T)$, let $\mathcal{A}(s,w)$ denote the set of all pairs (c,p) of consumption and premium plans on $[s,T]$, with

(a) $c(t) > 0$ for all $t \in [s,T]$;

(b) $Z(t) > \max\{0, -\frac{\phi}{1-\phi}c_b(t)\}$ for all $t \in [s,T)$;

(c) $W(T) > \max\{0, -\frac{\phi}{1-\phi}c_b(T)\}$.

The value function is then defined via $V(s,w) = \sup_{(c,p)\in\mathcal{A}(s,w)} J(s,w;c,p)$ and the optimization problem (2.4) can equivalently be written as $V(w) = V(0,w)$. Note that we do not need to define the set $\mathcal{A}(T,w)$ because the terminal boundary condition is known: $V(T,w) = B(T,w)$.

We now formulate the dynamic programming principle (DPP). For all $s < t$ in $[0,T]$ and w,

$$V(s,w) = \sup_{(c,p)\in\mathcal{A}(s,w)} \Big(\mathrm{e}^{-\beta(t-s)}\bar{F}(t,s)V(t,W(t)) \\ + \int_s^t \mathrm{e}^{-\beta(u-s)}\big(\bar{F}(u,s)U(c(u)) + f(u,s)B(u,Z(u))\big)\,\mathrm{d}u\Big).$$

The DPP gives rise to a numerical algorithm which we use to solve the optimization problem, see Algorithm 1 below.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Algorithm 1 Dynamic programming

Select a time grid $0 = t_0 < t_1 < \ldots < t_{N-1} < t_N = T$ with steps Δt;
Select (possibly time-dependent) nodes for the state variable $W(t)$: w_j, $j = 1, \ldots, J$;
Initialise $\hat{V}(t_N, w_j) = B(t_N, w_j)$ for $j = 1, \ldots, J$;
for $i = N-1, \ldots, 0$ **do**
    Interpolate/extrapolate $\hat{V}(t_{i+1}, w_j)$, $j = 1, \ldots, J$ to approximate $\hat{V}(t_{i+1}, w)$
    for any w (for $i = N-1$ the maturity condition is known: $\hat{V}(T, w) = B(T, w)$);
    Calculate $a_i^{(1)} = \int_{t_i}^{t_{i+1}} \mathrm{e}^{-\beta(s-t_i)} \bar{F}(s, t_i)\, \mathrm{d}s$, $a_i^{(2)} = \int_{t_i}^{t_{i+1}} \mathrm{e}^{-\beta(s-t_i)} f(s, t_i)\, \mathrm{d}s$
    and $a_i^{(3)} = \mathrm{e}^{-\beta(t_{i+1}-t_i)} \bar{F}(t_{i+1}, t_i)$, e.g. via numerical integration procedure;
    for $j = 1, \ldots, J$ **do**
        $\hat{V}(t_i, w_j) = \sup_{(c,p)} \left(a_i^{(1)} U(c) + a_i^{(2)} B(t_i, Z(t_i)) + a_i^{(3)} \hat{V}(t_{i+1}, \tilde{w}) \right)$,
        where $\tilde{w} = w_j(1 + r\Delta t) + (y(t_i) - c - p)\Delta t$;
    end for
end for
Calculate forward in time optimal trajectory
Set initial wealth, e.g. $W(0) = 0$
for $i = 0, \ldots, N-1$ **do**
    Find optimal controls
    $(c^*(t_i), p^*(t_i)) = \arg\sup_{(c,p)} \left(a_i^{(1)} U(c) + a_i^{(2)} B(t_i, Z(t_i)) + a_i^{(3)} \hat{V}(t_{i+1}, \tilde{w}) \right)$,
    where $\tilde{w} = W(t_i)(1 + r\Delta t) + (y(t_i) - c - p)\Delta t$;
    Find wealth at the next time: $W(t_{i+1}) = W(t_i)(1 + r\Delta t) + (y(t_i) - c^*(t_i) - p^*(t_i))\Delta t$.
end for
Calculated $(c^*(t_i), p^*(t_i))$ and $W(t_i)$ are optimal trajectories for consumption, insurance and wealth.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3. Solving Stochastic Climate-Economy Models: A Deep Least-Squares Monte Carlo Approach

The analysis of climate-economy policies is typically performed using Integrated Assessment Models (IAMs) that describe the complex interplay between the climate and the economy via deterministic equations. In order to account for stochastic shocks when finding optimal mitigation policies adapted to climate and economic variables that are evolving stochastically over time, a recursive dynamic programming implementation of integrated assessment models is required. This is a significantly harder computational problem to solve compared to the deterministic case. Seminal contributions to solving IAMs as optimal decision making problems in the presence of uncertainty include Kelly and Kolstad (1999), Kelly and Kolstad (2001), Leach (2007), Traeger (2014), and Cai and Lontzek (2019). All these studies are based on variants of the so-called dynamic integrated climate-economy (DICE) model extended to include stochastic shocks to the economy and climate. The DICE model is one of the three main IAMs (the other two being FUND and PAGE) used by the United States government to determine the social cost of carbon; see on Social Cost of Greenhouse Gases (2016). It has been regularly revised over the last three decades, with the first version dating back to Nordhaus (1992). It balances parsimony with realism and is well documented with all published model equations; in addition, its code is publicly available, which is an exception rather than the rule for IAMs. At the same time, it is important to note that IAMs, and the DICE model in particular, have significant limitations (in the model structure and model parameters), which have been criticized and debated in the literature (see the discussions in Ackerman et al. (2009); Pindyck (2017); Grubb et al. (2021); Weitzman (2011)). Despite the criticism, the DICE model has become the iconic typical reference point for climate-economy modelling, and is used in our study.

The original deterministic DICE model is solved as a global optimization problem using the General Algebraic Modeling Language (GAMS), a high-level programming language for mathematical modelling. Its stochastic extensions mentioned in the above-mentioned studies require implementations of recursive dynamic programming to find optimal climate policies under uncertainty (if required, the deterministic DICE model can be solved as a recursive dynamic programming problem, too). This is subject to the curse of dimensionality, and these studies are limited to only one or two stochastic variables. Even

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

in this case, computations take several million core hours on a modern supercomputer (see, for instance, Cai and Lontzek (2019)). Therefore, simulation methods are needed to handle models with many state variables and multiple shocks to reduce the computational burden.

The least-squares Monte Carlo (LSMC) method for solving multi-dimensional stochastic control problems has gained popularity in recent years due to its effectiveness in dealing with high dimensional problems and because it imposes fewer restrictions on the constraints and allows for flexibility in the dynamics of the underlying stochastic processes. The idea is based on simulating random paths of the underlying stochastic variables over time and replacing the conditional expectation of the value function in the Bellman backward recursive solution of the stochastic control problem with an empirical least-squares regression estimate. The transition density of the underlying process is not even required to be known in closed form; one just needs to be able to simulate the underlying processes. The LSMC method was originally developed in Longstaff and Schwartz (2001) and Tsitsiklis and Van Roy (2001). The convergence properties of this method are examined in Belomestny et al. (2010); Belomestny (2011), and Aïd et al. (2014). The LSMC method was originally developed for pricing American options where the state variables are not affected by the control. Later, an extension of the LSMC method with control randomisation was developed in Kharroubi et al. (2014) to handle endogenous state variables (i.e. state variables that are affected by controls). When applied to stochastic control problems that aim to optimize an expected utility, some further extensions are needed as proposed in Andréasson and Shevchenko (2022) and Andréasson and Shevchenko (2024) to achieve a stable and accurate solution.

In this chapter, we demonstrate how the LSMC method can be adapted to solve the recursive dynamic programming problem of stochastic IAMs. We exemplify this approach with an application to the DICE model with uncertainties in: (1) the equilibrium temperature sensitivity, (2) the damage function coefficient, (3) the growth rate of total factor productivity, (4) the growth rate of decarbonization, and (5) the equilibrium carbon concentration in the upper strata. These five uncertainties were identified in Nordhaus (2018) as being major sources of uncertainty for the evolution of climate-economic state variables. Typically, polynomial regression is used in LSMC to approximate the corresponding conditional expectations with respect to state variables and controls. However, for models such as the stochastic DICE model, this leads to the need of too many covariates and simulations, making the method not practical. To overcome this problem, we use deep neural network approximations for the required regressions and provide detailed explanations.

The DICE model is a deterministic approach that combines a Ramsey–Cass–Koopmans neoclassical model of economic growth (also known as the Ramsey growth model) with a simple climate model. It involves six state variables (economic capital; temperature in atmosphere and lower oceans; carbon concentration in atmosphere, upper and lower oceans) evolving deterministically in time, two control variables (savings and carbon

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

emission reduction rates) to be determined for each time period of the model, and several exogenous processes (e.g. population size and technology level). The uncertainty about the future of the climate and economy is then typically assessed by treating some model parameters as random variables (because we do not know the exact true value of the key parameters) using a Monte Carlo analysis (see Nordhaus (2018); Gillingham et al. (2015)).

Modelling aleatoric uncertainty owing to the stochastic nature of the state variables (i.e. owing to the process uncertainty that is present even if we know the model parameters exactly) requires the development and solution of the DICE model as a dynamic model of decision-making under uncertainty, where we calculate the optimal policy response under the assumption of continuing uncertainty throughout the time frame of the model. Few attempts have been made to extend the DICE model to incorporate stochasticity in the underlying state variables and solve it as a recursive dynamic programming problem. For example, Kelly and Kolstad (1999) and Leach (2007) formulated the DICE model with stochasticity in the temporal evolution of temperature, and solved this as a recursive dynamic programming problem. These studies are seminal contributions to the incorporation of uncertainty in the DICE model (although their numerical solution approach is difficult to extend to a higher dimensional space and time-frequency). Cai and Lontzek (2019) formulate DICE as a dynamic programming problem with a stochastic shock on the economy and climate. In addition, Traeger (2014) developed a reduced DICE model with a smaller number of state variables, whereas Lontzek et al. (2015) studied the impact of climate tipping points, and Shevchenko et al. (2022) considered the DICE model with discrete stochastic shocks to the economy. To our best knowledge, the only attempt to solve the stochastic DICE model using an LSMC-type approach is Ikefuji et al. (2020). Their study handles only one uncertainty at a time, and the setup of the regression type Monte Carlo algorithm omits the integration for the conditional expectation in the Bellman equation, assuming the randomness is known in the transition of state variables (in principle, in this case, the required integration can be performed by using deterministic quadrature methods, but this will be subject to the curse of dimensionality).

The primary contributions of this chapter are as follows:

(a) We introduce an efficient approach for modelling stochastic climate-economy models by combining the least-squares Monte Carlo method with deep learning techniques. It provides flexibility in handling various types of uncertainties, including both parametric and stochastic process uncertainties.

(b) We formulate a stochastic version of the DICE model using the sources of uncertainty as identified by Nordhaus (2018). Notably, it does not rely on discretizing the underlying probability distributions that is usually performed in Monte-Carlo type analyses for the sake of model tractability.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

(c) We perform comprehensive numerical experiments and discuss numerical techniques to significantly reduce the computational burden and address several peculiarities of the model. Moreover, we demonstrate how to perform uncertainty quantification (UQ) to understand how uncertainties in the model propagate and affect outputs (such as projections for the evolution of atmospheric temperature).

The chapter is organized as follows. Section 3.1 gives a description of the considered model. Section 3.2 describes the numerical method used to solve the model. Section 3.3 provides a comprehensive numerical study.

3.1. Model description

In this section, we present the DICE-2016R2 model as a classical example of a recursive climate-economy model. This version of the DICE model was used in Nordhaus (2018). It includes parameter uncertainties in equilibrium temperature sensitivity, the damage function coefficient and the equilibrium carbon concentration in the upper strata, as well as process uncertainties in the growth rate of total factor productivity and the growth rate of decarbonization.

The original deterministic DICE model seeks to find policies π that maximize a social welfare function, which models the discounted sum of population-weighted utility of per capita consumption:

$$V = \sup_{\pi} \sum_{t=0}^{\infty} \rho^t L_t \, u(c_t), \tag{3.1}$$

where ρ is a discount factor, L_t is the world population, c_t denotes per capita consumption, and the time index $t = 0, 1, \ldots$ corresponds to $\Delta = 5$-year steps. The policy $\pi = (\pi_t)_{t=0,1,\ldots}$ consists of two control variables, per capita consumption c_t and a carbon mitigation rate μ_t. The utility function u has constant elasticity with respect to per capita consumption, $u(c) = (u^{1-\alpha} - 1)/(1 - \alpha)$, with a risk-aversion parameter $\alpha \geq 0$ (the case $\alpha = 1$ corresponds to logarithmic utility).

The model features six state variables: economic capital K_t, the concentration of carbon in the atmosphere, the upper oceans, and the lower oceans, $M_t = (M_t^{\mathrm{AT}}, M_t^{\mathrm{UP}}, M_t^{\mathrm{LO}})^{\mathsf{T}}$, and the global mean temperature of the Earth's surface and the deep oceans, $T_t = (T_t^{\mathrm{AT}}, T_t^{\mathrm{LO}})^{\mathsf{T}}$. The evolution of the economic and geophysical sectors is governed by the dynamics described below.

The economic system: Gross output is modeled by a Cobb–Douglas production function of capital, labor, and technology, $Y_t = A_t K_t^{\gamma} L_t^{1-\gamma}$, where $\gamma \in (0, 1)$ and $1 - \gamma$ are the output elasticities of capital and labor, respectively. Here, A_t denotes *total factor productivity* (see Subsection 3.1.1), representing technological progress and efficiency improvements over time.

The DICE model incorporates economic damages from climate change, represented

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

by a damage function that is quadratic in the global mean surface temperature, $d_t = \pi_2 \times (T_t^{\mathrm{AT}})^2$, where π_2 is the *damage coefficient* (see Subsection 3.1.1). These damages can be mitigated by emission reduction, controlled by the policy μ_t. Reducing emissions incurs abatement costs Λ_t (see Table 3.1 for their specification).

Net output is then given by gross output reduced by damages and abatement costs, $Q_t = (1 - \Lambda_t)Y_t/(1 + d_t)$, and economic capital K_t evolves according to the following dynamics:

$$K_{t+1} = (1 - \delta_K)^{\Delta} K_t + \Delta \times (Q_t - C_t), \tag{3.2}$$

where C_t is total consumption, and δ_K is the rate of depreciation of economic capital.

The carbon cycle: The carbon cycle is modeled by three reservoirs, which follow the dynamics:

$$M_{t+1} = \Phi M_t + (\Delta \times \beta E_t, 0, 0)^{\mathsf{T}}, \tag{3.3}$$

where Φ is a coefficient matrix, E_t is total CO_2 emissions (in billions of tons per year), and β is the conversion factor of CO_2 mass into the equivalent mass of carbon. Emissions E_t are equal to uncontrolled industrial emissions, given by a level of *carbon intensity* (see Subsection 3.1.1) σ_t times gross output, reduced by the emission reduction rate μ_t, plus exogenous land-use emissions $\widetilde{E}_t$, i.e. $E_t = \sigma_t(1 - \mu_t)Y_t + \widetilde{E}_t$.

The temperature module: The relationship between greenhouse gas accumulation and increased radiative forcing is described by the function:

$$F_t = \eta \log_2 \left(M_t^{\mathrm{AT}} / \widetilde{M}^{\mathrm{AT}}\right) + \widetilde{F}_t,$$

which models the change in total radiative forcings from anthropogenic sources such as CO_2. It consists of exogenous forcings $\widetilde{F}_t$ plus forcings due to atmospheric concentrations of CO_2. Here, $\widetilde{M}^{\mathrm{AT}}$ is the preindustrial atmospheric carbon concentration. The evolution of global mean temperatures follows the dynamics:

$$T_{t+1} = \Psi T_t + \begin{pmatrix} \psi_1 F_{t+1} \\ 0 \end{pmatrix}, \tag{3.4}$$

where Ψ is a coefficient matrix, and ψ_1 is a model parameter. It is important to note that T_t is measured in terms of the absolute increase in temperature relative to the year 1900.

In DICE-2016R2, μ_t is assumed to be non-negative with an upper bound of 1, i.e. no negative industrial emissions are allowed. Table 3.1 summarizes the main coefficients of the model. Note that the number of time steps N is chosen such that $t = 0$ corresponds to the year 2015, while $t = N$ corresponds to the year 2500.

The social cost of carbon (SCC): The social cost of carbon (SCC) is a measure of the economic harm caused by emitting one additional ton of carbon dioxide (CO_2) into the atmosphere. It represents the present value of the damages associated with a marginal increase in CO_2 emissions in a given year. The SCC is typically expressed in

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

monetary terms (e.g. dollars per ton of CO_2) and is used to help policymakers evaluate the benefits of reducing emissions and compare the costs of different climate policies or regulatory actions aimed at mitigating climate change. The SCC can be calculated in the DICE model by:

$$SCC_t = -1000\beta \frac{\partial V_t / \partial M_t^{\mathrm{AT}}}{\partial V_t / \partial K_t}, \tag{3.5}$$

where V_t denotes the value function at time t, and β represents the CO_2 to carbon mass transformation coefficient.

Table 3.1. *Parameters for the base model.*

$N = 97$ time steps of $\Delta = 5$ years
$L_{t+1} = L_t(11.500/L_t)^{0.134}$, $L_0 = 7.403$ (in billions)
$A_{t+1} = A_t/(1 - g_A(t))$, $g_A(t+1) = g_A(t)\exp(-0.005\Delta)$, $A(0) = 5.115$, $g_A(0) = 0.076$
$\sigma_{t+1} = \sigma_t \exp(g_\sigma(t)\Delta)$, $g_\sigma(t+1) = g_\sigma(t)(1-0.001)^{\Delta}$, $\sigma_0 = \frac{35.85}{105.5(1-0.03)}$, $g_\sigma(0) = -0.0152$
$\Lambda_t = 550(1-0.025)^t/(1000\theta_2)\sigma_t\mu_t^{\theta_2}$
$\widetilde{E}_t = 2.6(1-0.115)^t$, $\widetilde{F}_t = (0.5 + t/34)\mathbb{1}_{t<17} + \mathbb{1}_{t\geq 17}$
$K_0 = 223$, $M_0^{\mathrm{AT}} = 851$, $M_0^{\mathrm{UP}} = 460$, $M_0^{\mathrm{LO}} = 1740$, $T_0^{\mathrm{AT}} = 0.85$, $T_0^{\mathrm{LO}} = 0.0068$
$\alpha = 1.45$, $\beta = 1/3.666$, $\gamma = 0.3$, $\rho = 0.015$, $\delta_K = 0.1$
$\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} & 0 \\ \phi_{21} & \phi_{22} & \phi_{23} \\ 0 & \phi_{32} & \phi_{33} \end{pmatrix}$, $\Psi = \begin{pmatrix} 1-\psi_1\psi_2-\psi_1\psi_3 & \psi_1\psi_3 \\ \psi_4 & 1-\psi_4 \end{pmatrix}$
$\phi_{21} = 0.12$, $\phi_{32} = 0.007$, $\phi_{11} = 1-\phi_{21}$, $\phi_{12} = \phi_{21}588/360$
$\phi_{22} = 1-\phi_{12}-\phi_{32}$, $\phi_{23} = \phi_{32}360/1720$, $\phi_{33} = 1-\phi_{23}$
$\psi_4 = 0.025$, $\psi_1 = 0.1005$, $\psi_3 = 0.088$, $\psi_2 = 3.6813/3.1$
$\eta = 3.6813$, $\widetilde{M}^{\mathrm{AT}} = 588$, $\pi_2 = 0.00236$, $\theta_2 = 2.6$

3.1.1. Modelling uncertainty

The dynamics presented in the DICE model so far are purely deterministic, assuming precise knowledge of the future evolution of all exogenous variables for centuries ahead. This approach is an unrealistic simplification. A reasonable way to address this issue is to introduce probabilistic distributions into the model to account for uncertainties about future outcomes. Here, we distinguish between two types of uncertainties: stochastic process uncertainty, and initial parameter uncertainty.

Stochastic process uncertainty refers to the uncertainty in the evolution of future trajectories of exogenous variables. A classical example from quantitative finance is Brownian motion, $B = (B_t)_{t\geq 0}$, modeled by $B_0 = 0$ and $B_{t+h} - B_t \sim \mathcal{N}(0, h)$ for $t \geq 0$ and $h \geq 0$, where $\mathcal{N}(0, h)$ denotes the normal distribution with expected value 0 and

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

variance h. Incorporating stochastic process uncertainties is challenging because the uncertainty propagates over time, increasing the volatility of the variable's distribution. The LSMC method we present below is highly sensitive to introduced volatility, making this incorporation a significant challenge that few contributions in the climate-economy literature have successfully addressed.

Initial parameter uncertainty refers to uncertainty about one or more parameters in the system that remain fixed over time. A common method to study this uncertainty is a perturbation analysis, where parameters are sampled, the model is solved, and the process is repeated. However, this approach does not accurately depict the model's evolution over time, as an agent in the model would consider overall outcome uncertainty, not individual instances of the uncertain parameter. Another related concept is Bayesian learning (Kelly and Kolstad, 1999), where the parameter distribution evolves over time as more information about the system is revealed. This type of uncertainty can be treated by the LSMC approach presented in this chapter, but we chose not to include this in the current study, leaving it for future work.

Identifying reasonable uncertainties to include in the model is challenging, as some uncertainties might be more significant than others. Advanced statistical analyses are required to make educated assumptions about probability distributions for the climate and economc system. Here, we incorporate five uncertainties into the DICE model, as identified by Nordhaus (2018). These include stochastic process uncertainties in the growth rates of total factor productivity A and the rate of decarbonization σ, as well as initial parameter uncertainties in the temperature-sensitivity coefficient, the damage coefficient, and the carbon cycle coefficient. We emphasize that our method is not limited to these specific uncertainties, and we now explain our choices in more detail.

Productivity growth. Assuming a Cobb-Douglas production function, the growth in total factor productivity A models the growth in output that is not explained by growth in inputs of labor and capital used in production. The DICE model assumes A evolves according to $A_{t+1} = A_t/(1 - g_A(t))$, where $g_A(t)$ is the deterministic growth rate which is specified in Table 3.1. Nordhaus (2018) assumes $g_A(0)$ is normally distributed with mean 0.076 and standard deviation 0.056. But in this case, using the dynamics for the growth rate, we can model $g_A(t)$ as normally distributed with mean $g_A(0)\exp(-0.005t\Delta)$ and standard deviation $0.056\exp(-0.005t\Delta)$. In order to remove extreme cases, we truncate this distribution at the mean $\pm$ two standard deviations. The evolution of A_t is shown in Figure 3.1.

The rate of decarbonization. Uncontrolled industrial CO_2 emissions are given by a level of carbon intensity, σ_t, times gross output. The DICE model assumes σ evolves according to $\sigma_{t+1} = \sigma_t \exp(g_\sigma(t)\Delta)$, with a deterministic growth rate $g_\sigma(t)$ which is specified in Table 3.1. Nordhaus (2018) assumes $g_\sigma(0)$ is normally distributed with mean -0.0152 and standard deviation 0.0032. We therefore model $g_\sigma(t)$ as normally distributed with mean $g_\sigma(0)(1 - 0.001)^{t\Delta}$ and standard deviation $0.0032(1 - 0.001)^{t\Delta}$, truncating the distribution at the mean $\pm$ two standard deviations in order to remove extreme cases.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

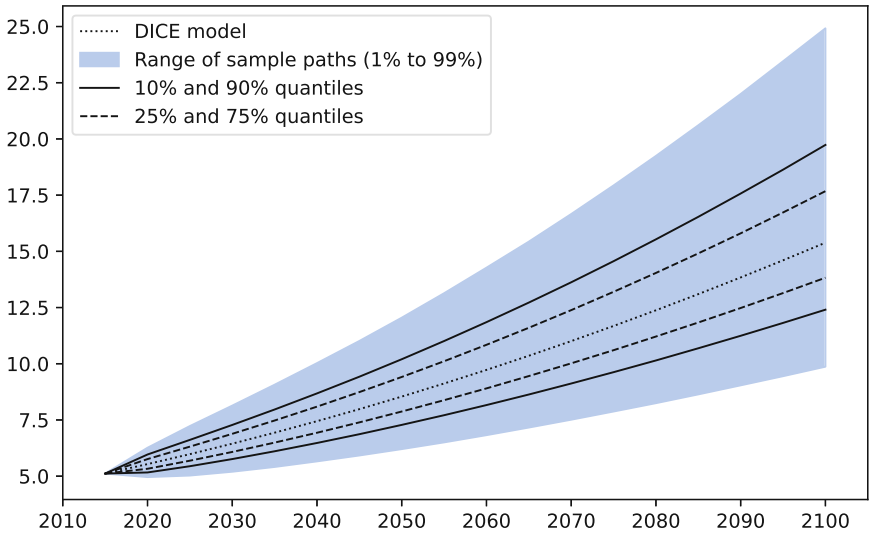

Figure 3.1. *Uncertain evolution of total factor productivity A under the assumption that the growth rate g_A is uncertain.*

The evolution of σ_t is shown in Figure 3.2.

Equilibrium temperature sensitivity (ETS). The equilibrium temperature sensitivity measures how much the Earth's surface will warm in response to a doubling of atmospheric CO_2. The DICE model assumes the ETS is equal to 3.1°C for an equilibrium CO_2 doubling. In Table 3.1, the ETS corresponds to the denominator in the definition of ψ_2. Nordhaus (2018) models the ETS as a log-normal distribution, $\exp(X)$ with $X \sim \mathcal{N}(1.1060, 0.2646^2)$. We do the same, truncating at the mean $\pm$ two standard deviations.

The damage function. The DICE model assumes climate-induced economic damages are a quadratic function of the increase in atmospheric temperature. It is modeled as a fractional loss of global output from greenhouse warming, $d(t) = \pi_2 \times (T_t^{\mathrm{AT}})^2$, where π_2 denotes a damage coefficient representing the severity of the economic impact of global warming. The DICE model assumes π_2 to be equal to 0.00236. Nordhaus (2018) models the π_2 by a normal distribution with mean 0.00236 and standard deviation 0.00118. We use the same distribution but truncate it at the mean minus one standard deviation (in order to avoid realizations of the damage coefficient that are too close to zero), and at the mean plus two standard deviations.

The carbon cycle. The carbon cycle coefficient models the equilibrium concentration of carbon in the biosphere and upper level of the oceans. The DICE model assumes it to be equal to 360 gigatonnes of carbon (GtC). In Table 3.1, it corresponds to the value 360 appearing in the definitions of ϕ_{12} and ϕ_{23}. Nordhaus (2018) models this coefficient

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

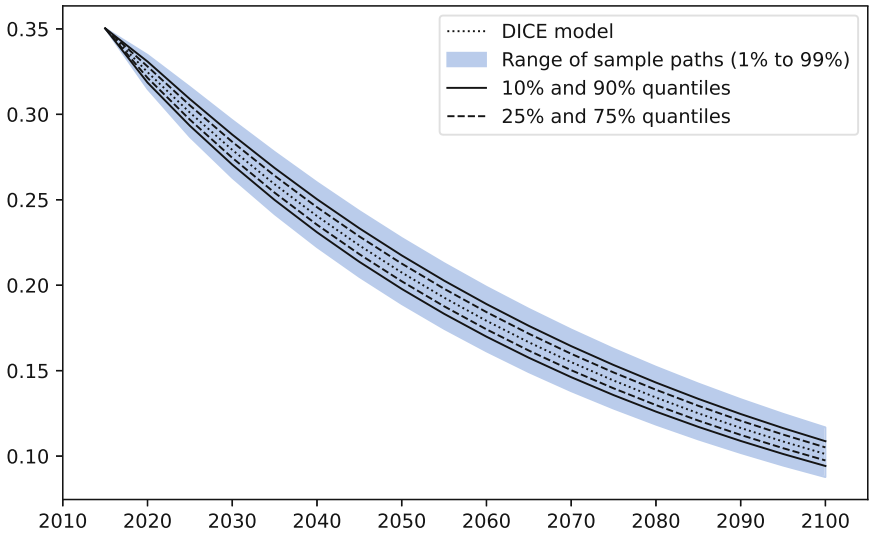

Figure 3.2. *Uncertain evolution of carbon intensity σ under the assumption that the growth rate g_σ is uncertain.*

as a log-normal distribution, $\exp(X)$ with $X \sim \mathcal{N}(5.8510, 0.2649^2)$ We do the same, truncating at the mean $\pm$ two standard deviations.

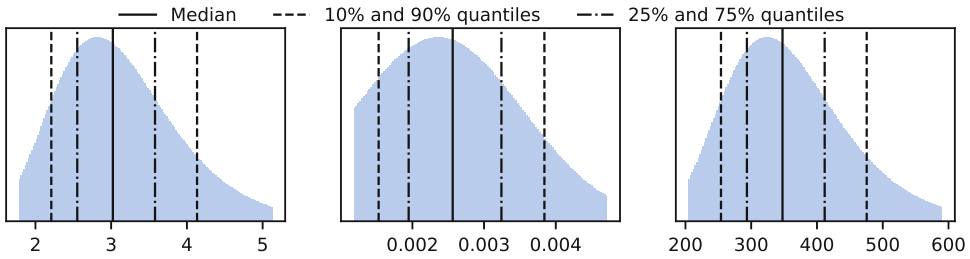

Figure 3.3. *Density plots of the parameter distributions of equilibrium temperature sensitivity (left panel), the damage coefficient (middle panel), and carbon cycle coefficient (right panel).*

Remark 3.1.

- Another type of uncertainty is *parametric uncertainty*, where the value of a coefficient can change over time as it is re-drawn at each point in time. This type of uncertainty lies between the stochastic process and the initial parameter uncertainty. Although we did not include it in our study, it is straightforward to incorporate and solve using our method.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

- Assuming $\pi_2 \sim \mathcal{N}(0.00236, 0.00118^2)$ implies a roughly 2.3% probability of π_2 being negative. This is a non-negligible scenario. Given that the DICE model aims to combine equations for the economy and climate, it is highly questionable to assume the damage coefficient could be below or just above zero. Moreover, the assumption of a log-normal distribution for the equilibrium temperature sensitivity and the carbon cycle coefficient also entails a non-negligible probability of those coefficients being close to zero.

  Nordhaus (2018) avoids this issue by discretizing the distributions, separating them into quintiles, and then calculating the expected values of the random variables within those quintiles. These expected values are taken as realizations of discrete uncertain variables, yielding sufficiently positive lowest realizations for the coefficients. Inspired by this approach, we also truncate the distributions of the random variables, however, without discretizing them. This avoids issues with too low damage coefficients and temperature sensitivities, as well as extreme growth rates for total factor productivity and carbon intensity.

3.2. The deep least-squares Monte Carlo method

The numerical solution of the model is achieved using the endogenous state least-squares Monte Carlo (LSMC) algorithm with control randomization, as introduced by Kharroubi et al. (2014) and adapted for expected utility optimal stochastic control problems by Andréasson and Shevchenko (2022). This method approximates the conditional expectation of the value function in the Bellman equation using regression with a quadratic loss function applied to the transformed value function. Typically, regression basis functions are ordinary polynomials of the state and control variables, usually up to the third order. However, with growing dimension of the state space, the usage of polynomials becomes increasingly difficult. Therefore, in our implementation, we use *deep neural networks* to approximate the regression predictor. To mitigate transformation bias in the regression estimate of the conditional expectation, we employ the smearing estimate as proposed by Andréasson and Shevchenko (2022). Below is a brief description of the LSMC algorithm.

Let $t = 0, 1, \ldots, N$ correspond to time points in the interval $[0, T]$. Consider the standard discrete dynamic programming problem with the objective to maximize the expected value of the utility-based total reward function

$$V_0(x) = \sup_{\pi} \mathbb{E}\left[\sum_{t=0}^{N-1} \rho^t R_t(X_t, \pi_t) + \rho^N R_N(X_N) \,\middle|\, X_0 = x; \pi \right], \tag{3.6}$$

where $\pi = (\pi_t)_{t=0,1,\ldots,N-1}$ is a control, $X = (X_t)_{t=0,1,\ldots,N}$ is a controlled state variable, R_N and R_t are reward functions, ρ is a time discount factor, and the expectation is

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

conditional on the initial state $X_0 = x$ and following the policy π. The evolution of the state variable is specified by a transition function $\mathcal{T}_t(\cdot)$ such that

$$X_{t+1} = \mathcal{T}_t(X_t, \pi_t, Z_t), \tag{3.7}$$

where $Z_0, Z_1, \ldots, Z_{N-1}$ are independent disturbance terms, i.e. the state of the next period depends on the current state's value, the current period's control decision, and the realisation of the disturbance term.

This problem can be solved using the backward recursion of the Bellman equation, starting from $V_N(x) = R_N(x)$ and then solving recursively:

$$V_t(x) = \sup_{\pi_t \in \mathcal{A}_t} \Big\{ R_t(x, \pi_t) + \mathbb{E}\big[\rho V_{t+1}(X_{t+1}) \,\big|\, X_t = x; \pi_t\big] \Big\}, \quad t = N-1, N-2, \ldots, 0, \tag{3.8}$$

where the expectation is conditional on the state $X_t = x$ and the policy π_t at time t. For further details on dynamic programming, we refer the interested reader to the excellent monograph by Fleming and Soner (2006) on the subject.

Using Equation (3.8), the optimal control can be found by solving:

$$\pi_t^*(x) = \arg\sup_{\pi_t \in \mathcal{A}_t} \Big\{ R_t(x, \pi_t) + \mathbb{E}\big[\rho V_{t+1}(X_{t+1}) \,\big|\, X_t = x; \pi_t\big] \Big\}. \tag{3.9}$$

Here, $\mathcal{A}_t$ denotes a set of admissible values of π_t, which may depend on x. When the number of state variables is more than three, it usually becomes computationally infeasible to use quadrature-based methods to evaluate the conditional expectation in (3.8), making simulation methods like LSMC preferable.

The LSMC method approximates the conditional expectation in Eq. (3.8):

$$\Phi_t(X_t, \pi_t) = \mathbb{E}\big[\rho V_{t+1}(X_{t+1}) \,\big|\, X_t; \pi_t\big] \tag{3.10}$$

using a regression scheme with the states X_t and randomized policies π_t as independent variables, and $\rho V_{t+1}(X_{t+1})$ as the response variable. The approximation function is denoted $\widehat{\Phi}_t$. The method is implemented in two stages:

(a) **Forward simulation:** For $t = 0, 1, \ldots, N-1$, the random state, control, disturbance variables as well as the transitioned state are simulated as X_t^m, π_t^m, Z_t^m, and $\widetilde{X}_{t+1}^m = \mathcal{T}_t(X_t^m, \pi_t^m, Z_t^m)$, $m = 1, 2, \ldots, M$, where π_t is sampled independently from X_t.

(b) **Backward recursion:** Starting from the boundary condition $V_N(x) = R_N(x)$, the optimal stochastic control problem in Eq. (3.6) is solved using the recursion in Eq. (3.8), as detailed in Algorithm 2.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

3.2.1. Transformation bias and heteroskedasticity

To mitigate challenges in approximating the value function due to the extreme curvature of utility functions, one can introduce a transformation $H(x)$ that mirrors the shape of the value function. In our implementation, we use:

$$H(x) = \frac{1}{1-\alpha} \mathrm{e}^{x(1-\alpha)}. \tag{3.11}$$

At each time $t < T$, the transformed value function is approximated using the least-squares regression:

$$H^{-1}(\rho V_{t+1}(X_{t+1}^m)) = \boldsymbol{f}_\theta(X_t^m, \pi_t^m) + \epsilon_t^m, \quad m = 1, 2, \ldots, M, \tag{3.12}$$

where ϵ_t^m, $m = 1, 2, \ldots, M$ are zero mean and independent error terms, $\{\boldsymbol{f}_\theta \colon \theta \in \Theta_t\}$ is a parametrized family of predictor functions, and H^{-1} the inverse of the transformation function. Then,

$$\Phi_t(X_t, \pi_t) = \int H(\boldsymbol{f}_\theta(X_t, \pi_t) + y) \, \mathrm{d}F_t(y), \tag{3.13}$$

where F_t is the distribution of the error term ϵ_t.

In the absence of a closed-form solution for the integral in Eq. (3.13), the empirical distribution of the residuals:

$$\widehat{\epsilon}_t^m = H^{-1}(\rho V_{t+1}(X_{t+1}^m)) - \boldsymbol{f}_\theta(X_t^m, \pi_t^m) \tag{3.14}$$

can be used to approximate this integral. Consequently, the estimate of $\Phi_t(X_t, \pi_t)$ becomes:

$$\hat{\Phi}_t(X_t, \pi_t) = \frac{1}{M} \sum_{m=1}^{M} H(\boldsymbol{f}_\theta(X_t, \pi_t) + \widehat{\epsilon}_t^m). \tag{3.15}$$

For the chosen transformation $H(x)$ in (3.11), Eq. (3.15) simplifies to:

$$\hat{\Phi}_t(X_t, \pi_t) = H(\boldsymbol{f}_\theta(X_t, \pi_t)) \frac{1}{M} \sum_{m=1}^{M} \mathrm{e}^{\widehat{\epsilon}_t^m (1-\alpha)}. \tag{3.16}$$

In Eq. (3.16), the mean of the transformed residuals does not depend on (X_t, π_t), simplifying the function evaluation of $\hat{\Phi}_t(X_t, \pi_t)$, as the mean can be precomputed and reused.

If heteroskedasticity is present in the regression with respect to the state and control variables, a method that accounts for heteroskedasticity is required. In this case, the conditional variance can be modelled as a function of covariates:

$$\mathrm{var}(\epsilon_t \mid X_t; \pi_t) = \boldsymbol{g}_\theta(X_t, \pi_t), \tag{3.17}$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

where $\{\boldsymbol{g}_\theta : \theta \in \widehat{\Theta}_t\}$ is another parametrized family of predictor functions. There are various standard methods to estimate $\boldsymbol{g}_\theta$ and the *smearing estimate with controlled heteroskedasticity* can then be used as discussed in Andréasson and Shevchenko (2022).

Remark 3.2.

- The method presented in Algorithm 2 is called the *regression surface approach*. A common alternative is the *realized value approach*, where the value function $V_t(x)$ in Eq. (3.8) is not computed by using the approximation of the conditional expectation (which was needed to find the optimal policy according to Eq. (3.9)), but rather by computing the discounted sum of rewards along one trajectory starting from the state x at time t. While promising greater numerical stability than the regression surface approach, the realized value approach requires calculating optimal decisions along the individual trajectories, which comes at a significant computational cost. For details on this approach, we refer to Andréasson and Shevchenko (2022) and references therein. Originally, we also implemented the realized value approach, however, we found that the regression surface approach provided a sufficiently accurate solution for the number of sample points chosen in our numerical study in Section 3.3.

- Another approach worth mentioning is the *regress later* LSMC method. Here, the value function is approximated directly rather than the conditional expectation: $V_{t+1}(x) \approx \boldsymbol{f}_\theta(x)$. Finding the optimal policy in (3.9) then requires the explicit calculation of the conditional expectation:

$$\mathbb{E}\big[\boldsymbol{f}_\theta(X_{t+1}) \,\big|\, X_t = x; \pi_t\big]$$

  either analytically or numerically with quadrature methods. However, as mentioned earlier, this approach becomes infeasible in the case of many simultaneous shocks due to the high dimensionality of the required integration.

3.2.2. Neural networks

Here, we choose for the parametrized family of functions $\{\boldsymbol{f}_\theta : \theta \in \Theta\}$ the class of deep neural networks. This algorithmically generated class of functions has found tremendous success in all fields of science. Over the years, it has been shown that neural networks can act as surrogate functions in many models, due to their far reaching approximation capabilities.

Theorems that establish approximations are referred to as *universal approximation theorems* (UAT); notable contributions include Cybenko (1989) and Hornik (1991). These theorems establish the topological density of sets of neural networks in various topological spaces. One speaks of the universal approximation property (Kratsios, 2021) of a class of neural networks. Unfortunately, these theorems are usually non-constructive. To

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

numerically find optimal neural networks, one typically combines backpropagation (see, for example, Rumerlhart et al. (1986)) with ideas from stochastic approximation (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Dvoretzky, 1956).

Assuming sufficient integrability, the conditional expectation in Eq. (3.10) is the orthogonal projection of $\rho V_{t+1}(X_{t+1})$ onto the subspace spanned by (X_t, π_t) in the space of square-integrable random variables. The universal approximation property of neural networks in this space (see, for instance, Hornik (1991, Theorem 1)) then justifies the approximation of $\Phi_t(X_t, \pi_t)$ by $\boldsymbol{f}_\theta(X_t, \pi_t)$ for a suitably chosen neural network $\boldsymbol{f}_\theta$.

Algorithm 2 LSMC (regression surface)

[Forward simulation]
for $t = 0$ **to** $N - 1$ **do**
  for $m = 1$ **to** M **do**
    sample X_t^m in the domain of its possible values ▷ State
    sample π_t^m in the domain of its possible values $\mathcal{A}_t$ ▷ Control
    sample Z_t^m from the distribution specified by the model ▷ Disturbance
    Compute $\widetilde{X}_{t+1}^m := \mathcal{T}_t(X_t^m, \pi_t^m, Z_t^m)$ ▷ Evolution of state
  end for
end for

[Backward recursion]
for $t = N$ **to** 0 **do**
  if $t = N$ **then**
    $\widehat{V}_t(\widetilde{X}_t) := R_N(\widetilde{X}_t)$
  else
    [Regression of transformed value function]
    $\widehat{\theta}_t := \arg\min_{\theta \in \Theta_t} \sum_{m=1}^{M} \left[\boldsymbol{f}_\theta(X_t^m, \pi_t^m) - H^{-1}(\rho \widehat{V}_{t+1}(\widetilde{X}_{t+1}^m)) \right]^2$
    Approximate conditional expectation $\widehat{\Phi}_t(X_t, \pi_t)$ using Eq. (3.15)
    for $m = 1$ **to** M **do**
      [Find optimal control]
      $\pi_t^*(\widetilde{X}_t^m) := \arg\sup_{\pi_t \in \mathcal{A}_t} \{ R_t(\widetilde{X}_t^m, \pi_t) + \widehat{\Phi}_t(\widetilde{X}_t^m, \pi_t) \}$
      [Update value function]
      $\widehat{V}_t(\widetilde{X}_t^m) := R_t(\widetilde{X}_t^m, \pi_t^*(\widetilde{X}_t^m)) + \widehat{\Phi}_t(\widetilde{X}_t^m, \pi_t^*(\widetilde{X}_t^m))$
    end for
  end if
end for

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

3.2.3. Uncertainty quantification

Uncertainty quantification (UQ) is a research field focused on understanding how uncertainties in model inputs, parameters, and other factors propagate through models to affect their outputs. This understanding is crucial for making informed decisions based on model predictions, particularly in complex systems where such decisions can have significant consequences. A key tool in UQ are Sobol' indices (Sobol', 2001), which are quantitative measures used in sensitivity analysis to apportion the variance of a model output to different input variables or combinations of input variables. By identifying the most important input variables and their interactions, Sobol' indices guide efforts to sort out the main factors which should be studied with care in complex models.

Sobol' indices provide a comprehensive view of how input variables and their interactions influence model outputs. They can be applied to any type of model, regardless of its complexity or the nature of its inputs and outputs. They are particularly valuable because they capture the effects of nonlinear interactions among input variables, which is critical for understanding complex systems. However, calculating Sobol' indices requires a large number of model evaluations, which can be computationally expensive for complex models. The accurate estimation of Sobol' indices also depends on efficient and adequate sampling of the input space.

Denote our stochastic DICE model by G, which maps model inputs X (such as the temperature-sensitivity coefficient) to model outputs $Y = G(X)$ (such as the projection of the global mean surface temperature in the year 2100). There are two main types of Sobol' indices.

First-order Sobol' indices S_i: These indices represent the contribution of a single input variable X_i to the output variance $\mathbb{V}(Y)$, ignoring interaction effects with other variables:

$$S_i = \frac{\mathbb{V}_{X_i}(\mathbb{E}_{X_{\sim i}}[Y \mid X_i])}{\mathbb{V}(Y)},$$

where $\mathbb{E}_{X_{\sim i}}[Y \mid X_i]$ denotes the conditional expectation of Y given X_i with respect to all inputs X except for X_i, and $\mathbb{V}_{X_i}(\cdot)$ denotes the variance with respect to X_i.

Total-order Sobol' indices S_{T_i}: These indices represent the contribution of an input variable to the output variance, including all interactions with other variables. They are defined as:

$$S_{T_i} = 1 - \frac{\mathbb{V}_{X_{\sim i}}(\mathbb{E}_{X_i}[Y \mid X_{\sim i}])}{\mathbb{V}(Y)},$$

where $\mathbb{E}_{X_i}[Y \mid X_{\sim i}]$ denotes the conditional expectation of Y with respect to X_i given all inputs X except for X_i, and $\mathbb{V}_{X_{\sim i}}(\cdot)$ denotes the variance with respect to all inputs X except for X_i.

First- and total-order Sobol' indices help determine which input variables are the most influential. Variables with high first-order indices have a strong direct effect, while those with high total-order indices are significant due to their interactions with other variables.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In Section 3.3, we will compute Sobol' indices for our five identified uncertainties and examine their effect on the most important model parameters. It is important to note that computing Sobol' indices in conjunction with the LSMC method involves solving the model with the backwards recursion (3.8) only once, and then generating a sufficiently large amount of forward trajectories to estimate the indices S_i and S_{T_i}.

3.2.4. Comparison with other methods and contributions

Jensen and Traeger (2014) analyze long-term economic growth uncertainty in a DICE-based assessment model with an infinite-horizon. They express uncertainty in terms of stochastic shocks to the growth rate of total factor productivity. The value function is approximated by Chebyshev polynomials, and the system is solved by value function iteration. The base model has only 3 physical state variables: capital K_t, atmospheric carbon M_t, and technology level A_t.

Nordhaus (2018) considers the same DICE model version as the one used in this chapter. Five uncertainties are identified, the same as those explained in Subsection 3.1.1. These uncertainties are treated as initial parameter uncertainties. The distributions are discretized to reduce the computational burden, thereby reducing the number of possible scenarios from an uncountably infinite amount to just a few thousand. A Monte-Carlo based parameter perturbation analysis is performed, where parameters are sampled, and then the corresponding deterministic version of the DICE model is solved. In contrast to Nordhaus (2018), we don't need to discretize the distributions, and we need to solve the model only once.

Cai and Lontzek (2019) also study a stochastic version of the DICE model, extending the deterministic 6-dimensional model to a stochastic 9-dimensional model. Two additional model dimensions are due to uncertainty in the evolution of total factor productivity, and one additional dimension is due to a stochastic tipping point process. The stochastic processes are discretized, and the resulting model is solved by value function iteration, where the value function is approximated by Chebychev polynomials. The model is solved with the Blue Waters supercomputer, using 110,688 cores in parallel, with computation times of up to 8 hours. While we do not include a tipping point process, our simulation based method drastically reduces the computational burden by solving our 11-dimensional (in contrast to the 9-dimensional version of Cai and Lontzek (2019)) model formulation on a 64 core machine within around 18 hours of computation time, depending on the amount of numerical precision that is required for the solutions. Expressed in terms of pure core hours (i.e. number of cores multiplied by total computing time), this amounts to a reduction in computing time of more than 99%.

Ikefuji et al. (2020) formulate a stochastic version of the DICE model considering one uncertainty at a time: a) uncertainty in the damage-abatement fraction, b) uncertainty in the damage parameter, c) uncertainty in the emissions-to-output ratio, and d) uncertainty in total factor productivity. These uncertainties are introduced by multiplying the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

corresponding deterministic DICE variables by stochastic disturbances. Thus, the number of state variables is the same as in the deterministic DICE (6). To the best of our knowledge, this is the only attempt to solve a stochastic version of the DICE model by using an LSMC type approach. They use least-squares regression with polynomial basis functions to approximate the value function, i.e. in the spirit of regress later LSMC. Here, we note that their regression type Monte Carlo algorithm setup omits the integration for the conditional expectation in the Bellman equation, assuming the random disturbance is known in the transition of state variables. In principle, the standard regress later LSMC can be implemented here to handle this type of uncertainty but it will be a subject of the curse of dimensionality in the case of more than one shock.

Friedl et al. (2023) present a method for solving integrated assessment models and performing uncertainty quantification. They exemplify their approach on a version of the DICE model with uncertainties in equilibrium temperature sensitivity (that contains a Bayesian learning component), and the damage function (represented by a stochastic tipping process). First, a deep neural network is trained to output, in particular, the optimal policies and value function at a given point in time, and then a Gaussian process-based model is trained to approximate quantities of interest such as the social cost of carbon in order to speed up the evaluation when calculating UQ metrics. In contrast to Friedl et al. (2023), our method approximates the conditional expectation rather than the policy functions, and then finds those by running an optimizer to solve Eq. (3.9). Approximating μ_t by a regression scheme is a challenging task, since the presence of the bounds (i.e. $0 \leq \mu_t \leq 1$) require a very careful choice of an appropriate regression scheme that can effectively interpolate the optimal policy, especially in the presence of extended periods when the policy is on the boundary. Our approach avoids this issue by finding the optimal policy through an optimizer which, once the conditional expectation has been approximated, can be performed with a high degree of numerical precision and speed. Moreover, the deep LSMC method requires performing a least-squares regression, where the loss function is the squared distance between the object of interest and the neural network prediction. This choice of loss function is significantly simpler, as it avoids the eleven individual components that enter the loss function based on an elaborate set of first-order conditions that are needed in the solution of Friedl et al. (2023). Finally, in contrast to Friedl et al. (2023), we find that there is no need to train an additional Gaussian process-based surrogate model to perform UQ for the quantities of interest (such as the social cost of carbon). Once the backward recursion (Eq. (3.8)) has been performed, a large amount of optimal trajectories for different realizations of uncertainties can be computed easily in order to perform UQ for the quantities of interest.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

3.3. Numerical study

In this section, we present the numerical results from applying the least-squares Monte Carlo method with transformation bias adjustment and neural network approximation of conditional expectations. For clarity, we emphazise that our state vector X_t consists of 11 variables: the six variables from the deterministic formulation of the DICE model (K_t, M_t, T_t), the two stochastic processes A_t and σ_t, as well as the three parameters discussed in Subsection 3.1.1 (temperature-sensitivity coefficient, damage coefficient and carbon cycle coefficient).

For the backward recursion and least-squares approximation of the value function, we use 2^{23} sample points in the 11-dimensional state space. Figure 3.6 is based on 5×10^5 forward trajectories, while the statistics reported in Table 3.2 are based on a sample of size 10^6. To find the optimal policies in (3.9), we use the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B). On a 64 core machine, it took between 9 hours (for 2^{22} samples) and 18 hours (for 2^{23} samples) to perform the backward recursion. Computing optimal forward trajectories then typically took around 15 minutes for 10^5 trajectories, and 1 hour for 5×10^5 trajectories.

The initial year for the version of DICE model used in Nordhaus (2018) is 2015, not 2020. For illustration purposes, during calculation of the optimal forward trajectories, we made the first policy decision (c_0, μ_0) deterministic and equal to the optimal decision in the deterministic version of the model. This amounts to starting the forward trajectories in the year 2020 with initial values that correspond to the optimal deterministic DICE states identified in Nordhaus (2018). Moreover, the original DICE model is formulated as an infinite-horizon control problem, see Eq. (3.1). However, our formulation of the LSMC method as discussed in Section 3.2 assumes a finite time horizon with N time steps ($N = 97$ in our case corresponding to $t = 0$ being the year 2015, and $t = N$ being the year 2500). Imposing a finite time horizon corresponds to a truncation of the problem, and one needs to choose an appropriate boundary reward function $R_N(x)$. Similarly, as in Cai and Lontzek (2019), our terminal reward function is computed by assuming that after 2500 the policies are fixed to $\mu_t = 1$, $c_t = 0.78$, and that the system evolves deterministically. The reward is then equal to the discounted sum of population-weighted utility of per-capita consumption from following the fixed policies for another 100 time steps. Due to discounting and the large amount of time steps, it is assumed that a different choice of boundary reward that far ahead in the future should have a negligible impact on results for the twenty-first century.

For approximating conditional expectations, we use deep feedforward neural networks with two hidden layers, each containing 32 hidden nodes with hyperbolic tangent (tanh) as activation function, and a linear readout in the output layer. Neural network training is performed using minibatch stochastic gradient descent with the Adam optimizer (Kingma and Ba, 2017). The initial learning rate is set to 10^{-3} and reduced to a minimum of 10^{-5} during training. Early stopping is implemented to avoid overfitting. During the backward

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

recursion, the trained neural network from one step (e.g. step $t+1$) is used as the initial neural network for the next step's training (step t), which reduces computation time.

For this version of the stochastic DICE model, the transition equation (3.7) can be separated into two transitions:

$$X_{t+1} = \widetilde{\mathcal{T}}_t(F(X_t, \pi_t), Z_t), \tag{3.18}$$

where the deterministic transition to the *post-decision* variable $\hat{X}_t = F(X_t, \pi_t)$ precedes the transition $X_{t+1} = \widetilde{\mathcal{T}}_t(\hat{X}_t, Z_t)$. This allows the conditional expectation in (3.8) to be simplified to:

$$\mathbb{E}\big[\rho V_{t+1}(X_{t+1}) \,\big|\, X_t; \pi_t\big] = \mathbb{E}\big[\rho V_{t+1}(X_{t+1}) \,\big|\, \hat{X}_t\big].$$

This method offers two main advantages: (1) dimension reduction in the covariates needed for the least-squares approximation of the conditional expectation, and (2) an increase in sampling efficiency by sampling only the post-decision states $\hat{X}_t$ rather than both X_t and π_t. Our method benefits significantly from using post-decision variables, and we found a notable improvement in numerical precision.

Economic capital K_t and total factor productivity A_t can grow quite rapidly over time, especially in scenarios where large growth in A_t meets a low consumption rate c_t. This poses an important numerical challenge, since an appropriate domain for sampling the state variables needs to be chosen with care. A popular solution to this issue, having been applied successfully in Jensen and Traeger (2014), is to normalize economic capital as follows. First, we re-write output to express it in terms of labor-augmenting technology: $Y_t = A_t K_t^{\gamma} L_t^{1-\gamma} = K_t^{\gamma}(\widetilde{A}_t L_t)^{1-\gamma}$, where $\widetilde{A}_t = A_t^{1/(1-\gamma)}$. Let $\widetilde{A}_t^{\text{det}}$ denote the deterministic trajectory of $\widetilde{A}_t$, where $g_A(t)$ is fixed to be equal to the expected value. Economic capital and output are then expressed in terms of units of effective labor: $k_t = K_t/(\widetilde{A}_t^{\text{det}} L_t)$, and $y_t = Y_t/(\widetilde{A}_t^{\text{det}} L_t)$. The state variable A_t can also be substituted by $\widetilde{A}_t$ and further normalized to $a_t = \widetilde{A}_t/\widetilde{A}_t^{\text{det}}$. In our simulations, we found that these normalization steps had a favorable impact on the precision of the numerical results.

Calculating the social cost of carbon (3.5) requires knowledge of partial derivatives of the value function with respect to atmospheric carbon concentration and economic capital. Since we do not have an analytic representation of the value function, we follow an approximation approach that was discussed in Traeger (2014), where Chebychev polynomials were used to approximate the value function. At each time t, we approximate the value function V_t by a neural network:

$$V_t(x) \approx \boldsymbol{g}_\theta(x), \tag{3.19}$$

for a suitable parameter vector θ. This approach strikes a balance between numerical precision and analytical tractability, applicable even in the presence of uncertainty. Note that the idea of approximating the value function by a neural network has already been carried out in Kelly and Kolstad (2001) where, however, the neural network approximation

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

was not used for computing the social cost of carbon.

The post-decision variables $\hat{X}_t$, representing the states X_t after decision π_t, have the same dimension as X_t. The sampling step in Algorithm 2 requires choosing an effective sampling distribution. One standard approach would be to put a high-dimensional grid of uniformly drawn points around the deterministic DICE solution. However, in order to improve numerical precision, low-discrepancy grids are favourable in order to keep the number of sample points needed to a reasonable amount. Latin hypercube sampling offers a more favourable distribution of grid points compared to uniform sampling. We chose to use Sobol' grid points (Sobol', 1967), which offer even higher numerical precision compared to Latin hypercube samples. Figure 3.4 shows the point distribution of a uniform and of a Sobol' grid for comparison. We found that using a low-discrepancy grid improved the numerical precision of the results.

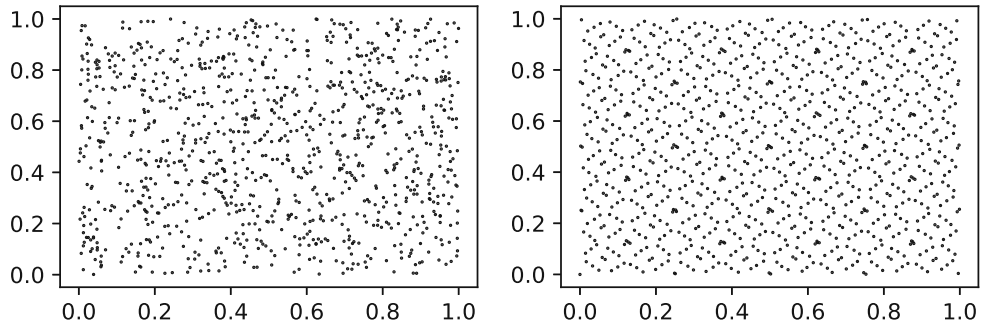

Figure 3.4. *Comparison of uniform grid (left panel) and low-discrepancy Sobol' grid (right panel). In both cases, 1024 points were drawn in 11 dimensions. The plots depict the point distributions from the 11-dimensional grid projected on the first two components.*

A major challenge in solving the model was to obtain stable estimates of the optimal emission mitigation rate μ_t. Estimating the optimal consumption rate c_t was straightforward, but estimating μ_t required very precise estimates in the least-squares approximation of the conditional expectation. Figure 3.5 offers a partial explanation. It illustrates a typical optimization surface when trying to find the optimal policies (c_t, μ_t) in Eq. (3.9), showing a steep curvature for c_t and a much flatter surface for μ_t, indicating the need for precise numerical approximations and small tolerance values in the optimizer. We see this issue as a consequence of the model setup. For example, a low carbon intensity σ_t for times after 2100 leads to low emissions and mitigation costs, resulting in an almost negligible effect of the mitigation rate on the value V_t. In order to resolve this issue, very precise numerical approximations of conditional expectations based on a large number of well-spaced sample points as well as small tolerance values in the optimizer for (c_t, μ_t) were required.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

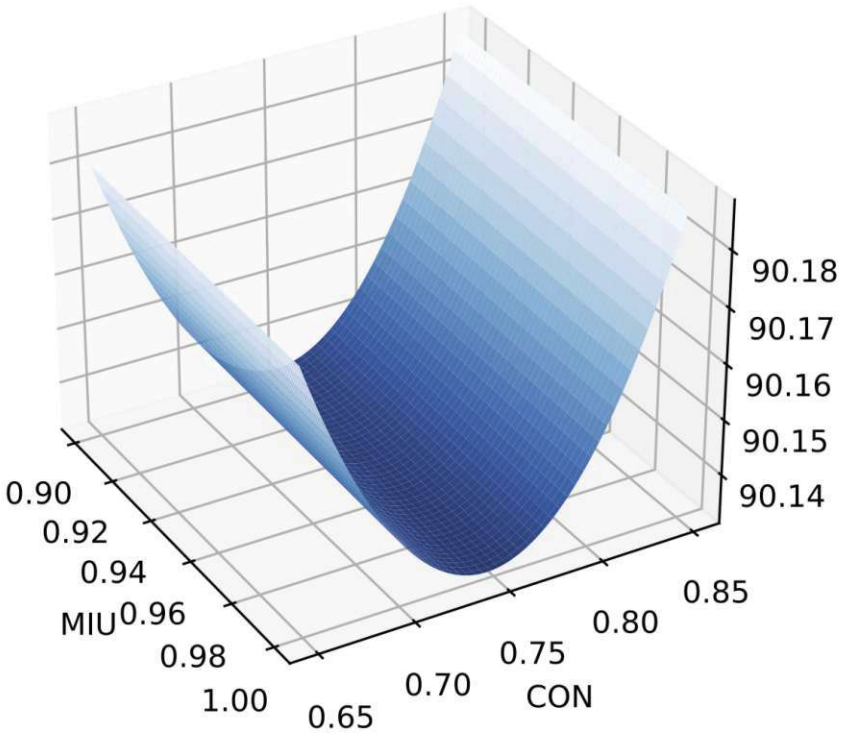

Figure 3.5. *Typical optimization surface over* (c_t, μ_t) *encountered during backward recursion.*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Each point x in the state space can be optimized independently in Eq. (3.9). In other words, when solving (3.9) over a high-dimensional grid in state space, the individual optimization steps for each grid point can be executed in parallel. This parallel optimization is implemented using Python's `multiprocessing` package over 64 cores, significantly reducing computation time and allowing for the usage of a reasonably large sample size without excessive computational costs.

Figure 3.6 presents the evolution of the six most important variables over time if the optimal strategy is used, based on 500,000 independently simulated trajectories. These six variables are the social cost of carbon SCC_t, the global mean surface temperature T_t^{AT}, the carbon concentration in the atmosphere M_t^{AT}, the emission mitigation rate μ_t, total CO_2 emissions E_t, and damages $\pi_2 \times (T_t^{\mathrm{AT}})^2$. The panels include the median trajectory (bold solid line), expected trajectory (dash-dotted line), the 25% and 75% quantiles (dashed lines), the 10% and 90% quantiles (solid lines) as well as the range of sampling paths between the 1% and 99% quantiles (shaded area). We can observe a significant amount of uncertainty in all variables. Most notably, a significant fraction of scenarios sees full mitigation (i.e. $\mu_t = 1$) well before the year 2100 in the optimal case, though the median trajectory is a bit below the full mitigation in 2100. We also observe that for temperature, the 1% quantile is approximately at 2.5°C, while the 99% quantile is approximately at 4.5°C. The SCC is about US$200 in 2100 under the median trajectory, and between $150 and $300 for the 10% and 90% quantiles. For all variables the median trajectory and deterministic DICE solution are virtually indistinguishable and very close to the expected trajectory.

Figure 3.7 shows the first- and total-order Sobol' indices for various model outputs in relation to the 5 sources of uncertainty which we considered in the model. The analyzed outputs are the social cost of carbon in 2020 (SCC), the mean surface temperature in the atmosphere in 2100 (TATM), the carbon concentration in the atmosphere in 2100 (MAT), output in 2100 (OUT), emissions in 2100 (EMI) as well as damages in 2100 (DAM). The first-order Sobol' indices (left panel) illustrate the individual contribution of each input to the variance of the outputs, while the total-order Sobol' indices (right panel) capture the overall contribution, including interactions with other inputs. Note that first-order indices do not sum up to 100%, as we have not taken into account higher order indices (second order, third order etc.).

From Figure 3.7, it is evident that output is predominantly impacted by total factor productivity, with both first-order and total-order indices close to 100%, indicating a strong direct influence. In contrast, the overall impact of the carbon intensity is negligible, with the indices being below 1% throughout. Uncertainty in σ_t could potentially be excluded to simplify the model without sacrificing accuracy. The temperature-sensitivity and damage coefficients exhibit high indices across all remaining outputs, implying their large influence on the model outputs. Both of these coefficients moreover show a significant difference between their first-order and total-order indices for emissions, suggesting substantial interaction effects with other inputs. Notably, the almost negligible

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

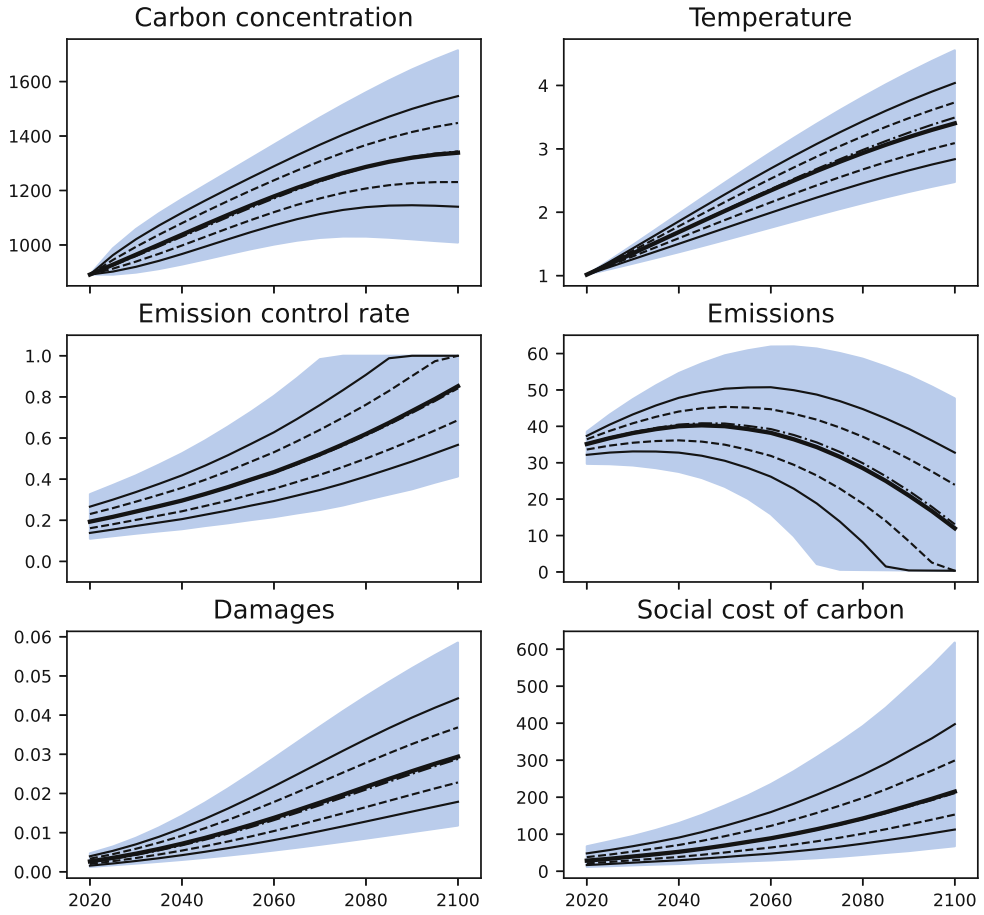

Figure 3.6. *Evolution of the six most important variables over time.*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

first- and total-order indices for the carbon cycle coefficient with respect to emissions is contrasted by significant indices for damages, as well as atmospheric temperatures and carbon concentrations. Finally, we observe that uncertainty in the social cost of carbon in 2020 is largely due to temperature-sensitivity and damage coefficients. This does not come as a surprise, as the uncertainty in A_t and σ_t propagates through time and is therefore not very pronounced in the year 2020 (compared to, for instance, the year 2100).

Overall, Figure 3.7 highlights that:

(a) Productivity has a strong influence on output, but neither on damages nor on temperatures.

(b) The carbon intensity has a completely negligible impact on the model.

(c) The temperature-sensitivity and damage coefficients have very strong impacts on the model.

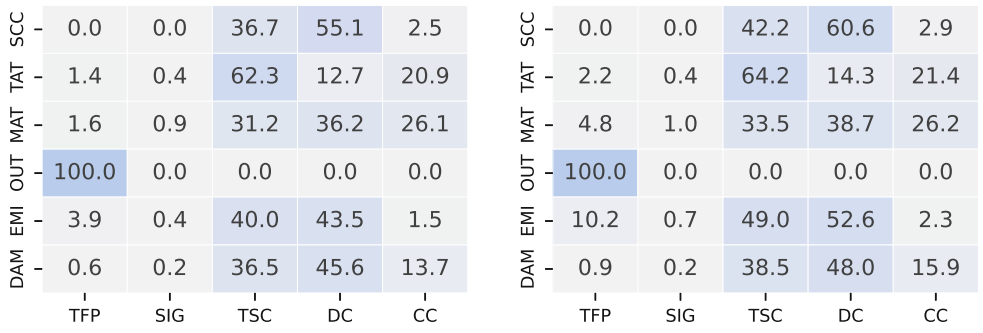

Figure 3.7. *First-order (left) and total-order (right) Sobol' indices for various model outputs with respect to uncertainty in total factor productivity (TFP), carbon intensity (SIG), temperature-sensitivity coefficient (TSC), damage coefficient (DC) and carbon cycle coefficient (CC).*

Figure 3.8 shows the evolution of first-order Sobol' indices for our main variables over time, up to the year 2150. It highlights the fact that the impact of the uncertain variables on the outputs changes over time. Most notably, the changes appear not to follow a linear pattern, especially when looking at emissions. There, the impact of total factor productivity A_t peaks around the year 2035, but declines rapidly afterwards. In contrast, the impact of A_t on the social cost of carbon gradually rises from 0% in the year 2020, to around 25% in the year 2150. This does not come as a surprise, as it highlights the effect of the large initial uncertainty about parameters such as the temperature-sensitivity and damage coefficients, which combines with a negligible initial uncertainty in total factor

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

productivity that grows over time. Another interesting effect that can be observed is that the total sum of all first-order indices declines for emissions from above 95% in the year 2020 to slightly below 40% in the year 2150. This motivates the insight that the impact due to interactions between the uncertain variables grows over time.

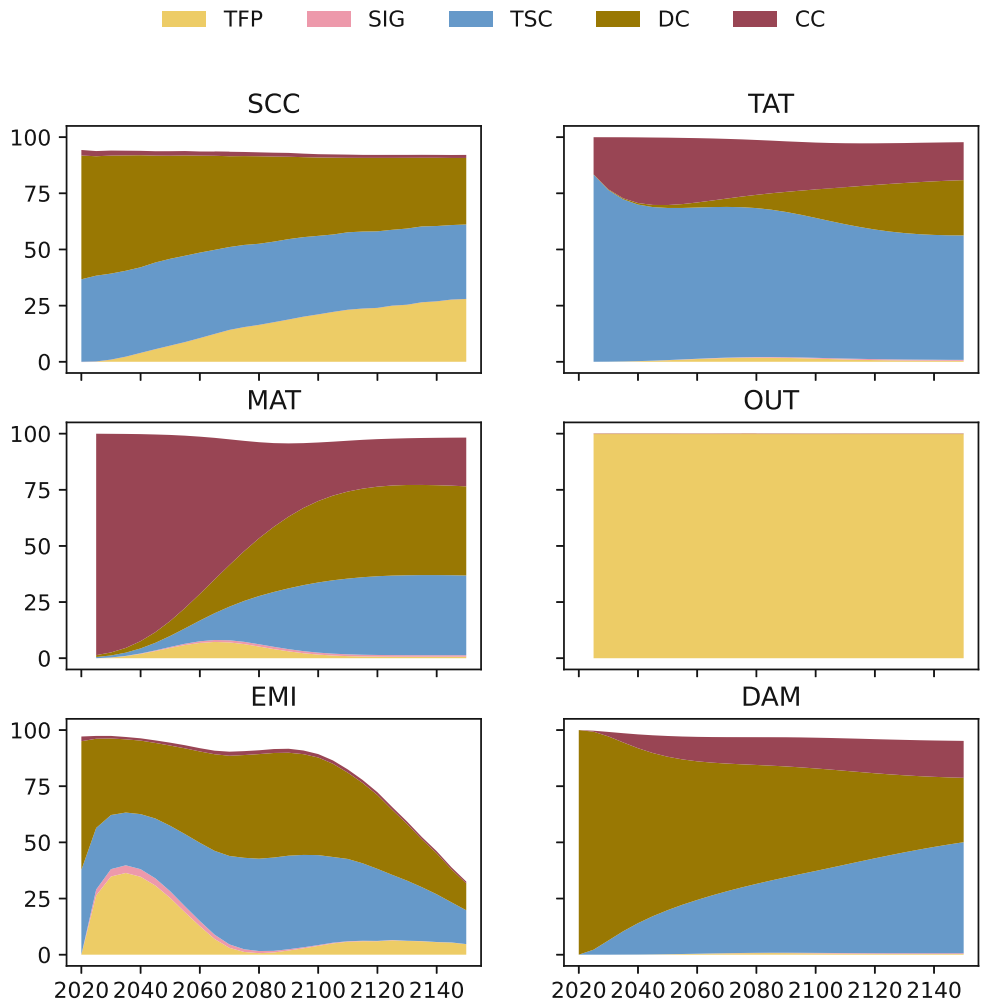

Figure 3.8. *First-order Sobol' indices for main variables over time.*

Table 3.2 shows the key statistics for the major variables. In terms of the coefficient of variation (CV), we can observe the highest degree of uncertainty in emissions, followed by the social cost of carbon, damages, and output. Most importantly, the interquartile range (IQR) of 0.64°C for temperature and 1.4% for damages highlights the importance of considering the notable variations in projections due to the presence of uncertainty.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Moreover, we can re-confirm the presence of noticeable differences between the mean, median and best guess values for some variables, which is in line with the observations of Nordhaus (2018). Differences between the mean and median values hint at the presence of skewness in the distribution of the variables, which can also be visually confirmed from Figure 3.6. Finally, differences between the best guess estimates and the mean and median values show that in some cases, the best guess provides a reasonable approximation of the complex dynamics, whereas in other cases it does not, which again highlights the importance of explicitly including stochastic dynamics into climate-economy models.

Table 3.2. *Statistics for major variables.*

Variable	**Mean**	**BG**	**Median**	**SD**	**IQR**	**CV**
Social cost of carbon, 2020	30.9	28.3	28.7	12.5	16.7	0.40
Temperature, 2100 (°C)	3.42	3.49	3.40	0.46	0.64	0.13
Carbon concentration, 2100 (ppm)	1,342	1,344	1,339	156	217	0.12
World output, 2100 (trillions, 2015$)	833.6	795.9	811.2	203.6	271.9	0.24
Emissions, 2100	14.0	13.1	12.0	13.3	23.6	0.95
Damages, 2100 (percent output)	3.0	2.9	2.9	1.0	1.4	0.34

SD, IQR and CV refer to standard deviation, interquartile range and coefficient of variation, respectively. BG refers to best guess, which is the value calculated along the expected trajectory, assuming that uncertainties are set to their respective means.

Conclusions

Climate-economy models are essential tools for informed decision-making, risk management, and strategic planning in the face of climate change. These models provide a structured framework for analyzing the economic implications of climate policies and developing sustainable solutions to mitigate and adapt to climate change impacts. Incorporating stochastic models into climate-economy analyses is crucial for capturing the full spectrum of uncertainties, improving risk assessment, designing resilient policies, and enhancing the overall robustness and reliability of the models and their predictions. However, the complexity of capturing the intricate, multifaceted, and probabilistic nature of climate and economic systems, coupled with the computational challenges of handling large-scale, high-dimensional, and stochastic models, poses significant challenges in deriving efficient solutions in the presence of uncertainty.

We present an advanced approach to modelling recursive stochastic climate-economy models using a deep least-squares Monte Carlo (LSMC) method. The method's flexibility allows for the application to various types of uncertainties, including parametric and stochastic process uncertainties. The integration of deep neural networks enables the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

handling of high-dimensional models in a tractable manner and within a reasonable computational budget, thus making stochastic climate-economy models more accessible to researchers and policymakers. The methodology and findings presented here provide a solid foundation for future work in this vital area of research.

Future research should explore the incorporation of Bayesian learning mechanisms to update probabilities as more information becomes available over time. Since our approach can manage high-dimensional stochastic shocks, a natural next step is to study the impact of multi-dimensional probability distributions whose marginals are correlated. Additionally, we aim to apply our method to the study of climate tipping points as well as the Regional Integrated model of Climate Change and the Economy (RICE) of Nordhaus and Yang (1996). These future steps could further refine the model's predictions and enhance its policy relevance.

It is important to note that IAMs, and the DICE model in particular, have limitations in the model structure and model parameters which are debated in the literature, see e.g. discussions in Pindyck (2017). The incorporation of uncertainties into these models is an important improvement. Our approach demonstrates significant advancements in modelling and solving complex stochastic climate-economy models. By capturing a wide range of uncertainties and using advanced computational techniques, we contribute to the development of more robust and reliable tools for climate policy analysis. The continued evolution of these models will be critical in supporting effective and sustainable climate action in the years to come, and the deep least-squares Monte Carlo method provides a useful tool to solve stochastic climate-economy models.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Part III.

Algorithmic Strategies

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

4. Reinsurance with Neural Networks

In 1903 Filip Lundberg suggested to model the surplus of an insurance company by a constant drift minus a compound Poisson process with iid almost positive jumps (Lundberg, 1903). The drift is interpreted as the premium rate, the jumps represent the claim sizes, whereas the Poisson process counts the claims. This setting, widely known as the classical risk model or Cramér–Lundberg model, gives a clear but simplified picture of the insurer's balance.

For an insurance company, claims are not the only source of uncertainty. For example, the increase or decrease in the number of customers (see Braunsteins and Mandjes (2023)) or a random interest rate (see Eisenberg (2015)) will impact the risk process. Also, reputational considerations, investments in positively correlated financial markets, and correlations between different business branches and collectives of insured will play a considerable role. Recently, one started to consider models involving a dependence between the actuarial business and financial markets offering investment possibilities, see, for instance, Ceci et al. (2022), Leimcke (2020) and references therein.

An important modification of the classical risk model is due to Gerber (1970), who suggested to include an additional source of uncertainty – a Brownian motion. This new, perturbed classical risk model can better account for reality while still being a one-dimensional Markov process. The latter property makes the perturbed process a popular model for the surplus of an insurance company, see, for instance, Dufresne and Gerber (1991), Tsai (2001), Cheung and Liu (2023) and references therein.

However, by adding an additional source of uncertainty, in many optimisation settings, the calculations and proofs become much more complicated, resulting in the use of the viscosity approach, see for instance Eisenberg (2015). The viscosity approach allows to find the optimal strategy numerically, since the corresponding Hamilton–Jacobi–Bellman equation can be tackled using the finite difference method. To avoid this approach, one may discretize the surplus process, thereby allowing controls only at discrete time points. In particular, for a finite time horizon, this method has the advantage that all control strategies can be written in feedback form, i.e. depending on the finite number of the observed state values.

In actuarial control theory, after the model for the surplus has been chosen, the question arises which risk measure will be considered as a target to optimize. The most famous and extensively studied risk measure, suggested by Lundberg (1903), is the ruin probability, i.e. the probability that the surplus becomes negative in finite time. A vast number of results have appeared over the last century concerning the minimisation of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the ruin probability in different settings. We refer the interested reader to Schmidli (2008), Asmussen and Albrecher (2010) and references therein. Alternative risk measures that have been considered over the last decades include in particular expected dividends, expected utility of terminal wealth, and expected capital injections, see, for instance, Albrecher et al. (2017), Avanzi (2009), Albrecher and Thonhauser (2009) and references therein.

Measuring the utility of the terminal wealth was first suggested by Borch (1961) and had since become one of the most important risk indicators in insurance mathematics. The wealth at some finite time T – for instance, the time of a regulatory check – can provide useful clues about the company's wellbeing. Including the ruin probability into the target functional has been considered for instance in Hipp (2018), where the main target is to maximize dividends, or in Thonhauser and Albrecher (2007), where the time value of ruin is taken into account. However, the multi-objective goal of simultaneously optimizing a risk measure (such as, in our case, the expected utility of terminal wealth) along with the ruin probability remains largely unexplored.

In the present manuscript, we look at a very general extension of the perturbed risk model that has not been considered before. The surplus of an insurance company is modelled by a jump process perturbed by a general diffusion (not necessarily a Brownian motion) on an interval $[0, T]$ with a deterministic time horizon T. This implies that the problem we consider is 3-dimensional and depends on the time to maturity, on the state of the jump process and on the state of the diffusion. The functional to maximize is given by the expected utility of terminal wealth perturbed by a modified Gerber–Shiu (Gerber and Shiu, 1998) penalty function. It is optimized over the class of reinsurance policies, which are arguably the most popular type of controls in the literature (with investments, dividends and capital injections being common alternatives). A substantial body of literature exists on utility maximization and ruin minimization with reinsurance strategies, with notable contributions including Schmidli (2002), Promislow and Young (2005), Bai and Guo (2008), Schmidli (2001) and Taksar and Markussen (2003).

The role of the penalty function here is twofold. It rewards the insurer if the surplus remains positive at all times (i.e. in the case of no ruin), while a negative surplus is penalised. In addition, one can opt for different weights for the expected utility and for the penalty function, depending on the individual preferences of the insurer. It means, we allow the insurer to prioritize their immediate needs: higher utility of the terminal surplus with higher risk, or a more safer play.

As it seems unlikely that this problem can be solved explicitly, we seek for the optimal strategy in the class of feedback controls using machine learning techniques. More specifically, we use neural networks. Neural networks have become a popular tool in actuarial risk management, having been applied to mortality modelling, claims reserving, non-life insurance pricing and telematics. For a survey on recent advances of artificial intelligence in actuarial science, see Richman (2021) and references therein.

The task of finding optimal reinsurance (and dividend) strategies with neural networks

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

has been studied in Cheng et al. (2020) and Jin et al. (2021). There, the authors develop a hybrid Markov chain approximation-based iterative deep learning algorithm to maximize expected dividends under consideration of the time value of ruin. In contrast to Cheng et al. (2020) and Jin et al. (2021), we include the ruin probability explicitly as a risk measure (besides the expected utility of terminal wealth) that is optimized. Moreover, we formulate our optimization problem as an empirical risk minimization problem, which can be solved efficiently by stochastic gradient descent methods, even in highly complex model settings.

The primary contributions of this chapter are as follows:

(a) We introduce a novel framework for optimizing reinsurance strategies using deep learning techniques in order to maximize a target functional comprising a utility function penalized by an extended Gerber–Shiu function. The proposed method allows the insurer to balance between maximizing the expected utility of terminal wealth, and minimizing the probability of ruin.

(b) By drawing connections to binary classification problems and surrogate loss functions, we demonstrate how the optimization problem can be solved by empirical risk minimization, a method that, when combined with stochastic gradient descent, is particularly useful for optimizing neural networks.

(c) We illustrate our proposed methodology by a numerical example, where the surplus process is given by a Cramér–Lundberg model perturbed by a mean-reverting Ornstein–Uhlenbeck process. Our findings demonstrate the effectiveness of our method in finding optimal reinsurance strategies, and highlight the large scope of the approach.

The chapter is organized as follows. Section 4.1 gives a mathematical description of the considered model. Section 4.2 introduces algorithmic reinsurance strategies. Section 4.3 exemplifies our approach with numerical experiments.

4.1. Model description

We consider an insurer with a deterministic finite planning horizon $T > 0$. The insurer manages a portfolio of risks that generate premium payments. To mitigate potential large financial losses from unexpectedly high claim frequencies or sizes, the insurer can enter into reinsurance agreements. For a given number of time steps $n \in \mathbb{N}$, these agreements are re-negotiated at time points $(t_i)_{i=0}^n$, $0 = t_0 < t_1 < \ldots < t_{n-1} < T$ for the coverage period $(t_i, t_{i+1}]$, where we set $t_n = T$.

Our study is based on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The flow of information is modeled by an $\mathbb{R}^r$-valued stochastic process Y, where $r \in \mathbb{N}$ is a fixed dimension. The process Y induces a filtration $\mathbb{F} = (\mathcal{F}_i)_{i=0}^n$ of $\mathcal{F}$, allowing for the possibility that $\mathcal{F}_n \neq \mathcal{F}$. That

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

is, there might be some information that does not reveal itself even at maturity T. We assume that $\mathbb{R}^r$ is endowed with the Borel-σ-algebra $\mathcal{B}_{\mathbb{R}^r}$. All stochastic processes are indexed via the discrete time points $(t_i)_{i=0}^n$.

Let $p = (p_i)_{i=0}^n$ denote the $\mathbb{F}$-adapted premium process. The payment p_i ensures insurance coverage over the period $(t_i, t_{i+1}]$. As in our numerical illustrations in Section 4.3, p_i might be computed according to the expected value principle with a positive safety loading. Alternatively, p_i can be calculated using any other known premium calculation principle such as, for instance, the standard deviation or zero-utility principle. The premium may also depend on the number and size of the previously occurred claims. However, for the sake of clarity, we concentrate on the expected value principle in order to better explain the features of our model, and leave further extensions to future research. Since the time horizon T is assumed to be finite, we may assume, without loss of generality, that $p_n = 0$.

Inspired by the Cramér–Lundberg model, let $N = (N_i)_{i=0}^n$ be an $\mathbb{F}$-adapted, $\mathbb{N}_0$-valued and increasing process with $N_0 = 0$, and N_i represents the number of claims up to time t_i. The $\mathbb{R}_+$-valued insurance claims are denoted by $(Z_i)_{i\geq 1}$. We also consider a real-valued, $\mathbb{F}$-adapted process $L = (L_i)_{i=0}^n$ which represents random fluctuations, such as small claims and variations in premium income.

Remark 4.1. A distinguishing feature is that p, N, $(Z_i)_{i\geq 1}$ and L are not assumed to be independent.

The reinsurance agreement is characterized by a reinsurance strategy $b = (b_i)_{i=0}^n$. For illustrative purposes, we assume the agreement to be proportional, that is, b is an $\mathbb{F}$-adapted process with values in $[0, 1]$. Here, b_i represents the retention level, and $(1 - b_i)$ is the proportion of claims covered by the reinsurer during the period $(t_i, t_{i+1}]$. For notational convenience, we set $b_n = 1$. The reinsurance premium is given by the process $c = (c_i)_{i=0}^n$. We assume that $c_i = f_i(b_i)$ for some continuous cost function $f_i\colon [0, 1] \to \mathbb{R}$.

For each reinsurance strategy b, the surplus process $X^b = (X_i^b)_{i=0}^n$ is defined by $X_0^b = x$, and

$$X_{i+1}^b = X_i^b + p_i - c_i + L_i - b_i \sum_{j=N_i+1}^{N_{i+1}} Z_j, \quad i = 0, 1, \ldots, n-1, \tag{4.1}$$

where $x \in \mathbb{R}$ is the initial capital. The surplus process without reinsurance is denoted $X = (X_i)_{i=0}^n$.

The insurer's preference is described by a continuous utility function $u\colon \mathbb{R} \to \mathbb{R}$. Here, we assume that the utility function is chosen such that $\mathbb{E}[u(X_n)]$ is finite. For example, in Section 4.3 we will choose an exponential utility function in conjunction with exponentially distributed claims. If one were to choose Pareto-distributed claims – a popular choice in the literature – then another utility function is required due to the heavy tails of the Pareto distribution.

Definition 4.2. A reinsurance strategy b is called *admissible*, if $\mathbb{E}[u(X_n^b)]$ is finite. The

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

set of all admissible reinsurance strategies is denoted by $\mathcal{A}$.

In our model, the insurer aims to optimize the expected utility of terminal wealth while considering the probability of ruin across all admissible reinsurance strategies. This is done by incorporating a penalty term whose strength is expressed by a parameter $\beta \in [0,1]$. That way, the target is to solve the following optimization problem

$$\sup_{b\in\mathcal{A}} \Big\{ \beta \cdot \mathbb{E}[u(X_n^b)] - (1-\beta) \cdot \mathbb{P}(\min_{0\le i\le n} X_i^b < 0) \Big\}. \tag{4.2}$$

Remark 4.3. In (4.2), the two boundary cases $\beta = 1$ and $\beta = 0$ correspond to the pure expected utility maximization and ruin probability minimization problems, respectively. More generally, the convex combination in (4.2), is a linear scalarization of a multi-objective optimization problem, i.e. the original problem with two objectives – the expected utility of terminal wealth and the ruin probability – is transformed into a single-objective optimization problem. Our approach can also accommodate other objectives alongside those discussed in this chapter; we leave the investigation of other objectives to future research. We also refer to Feinstein and Rudloff (2024), where the efficient frontier, i.e. the set of all optimal solutions for a multi-objective optimization problem, is approximated via neural networks.

Our problem formulation is kept very general, so that various notable models can be considered as special cases. For example, N might be a self-exciting process observed at discrete time points. As in Section 4.3, L could be (the discretization of) an Ornstein–Uhlenbeck process. Additionally, let us note that the formulation of this section extends naturally to the multi-dimensional case with, for instance, multiple correlated business lines.

We aim to solve the optimization problem (4.2) using empirical risk minimization and stochastic approximation, a classical concept in statistical learning theory which can be outlined as follows. Given a set of training data points $\{(Y^j, Z^j) \,|\, j = 1, 2, \dots, m\}$ that are independent and identically distributed, and a parametrized family of predictor functions $\{f_\theta \colon \theta \in \Theta\}$, the goal is to minimize the empirical loss,

$$\theta^* = \arg\min_{\theta\in\Theta} \frac{1}{m} \sum_{j=1}^{m} \ell(f_\theta(Y^j), Z^j),$$

where ℓ denotes a loss function. If $\ell(f_\theta(Y^j), Z^j)$ is differentiable with respect to θ, the gradient descent method can be used to minimize the empirical loss starting from an initial guess θ_0 and iteratively updating

$$\theta_{k+1} = \theta_k - \frac{\eta}{m} \sum_{j=1}^{m} \nabla_{\theta_k} \ell(f_{\theta_k}(Y^j), Z^j), \quad k = 1, 2, \dots,$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

until a pre-determined termination criterion is reached. Here, $\eta > 0$ denotes a learning rate, and $\nabla_\theta \ell(f_\theta(Y^j), Z^j)$ denotes the gradient of $\ell(f_\theta(Y^j), Z^j)$ with respect to θ.

For very large sample sizes, where computing the average gradient is numerically expensive, *stochastic gradient descent* (SGD) is a more efficient alternative, updating parameters based on individual gradients $\nabla_\theta \ell(f_\theta(Y^j), Z^j)$. However, due to the noisy nature of single-gradient updates, *mini-batch* stochastic gradient descent is often preferred. This method updates parameters based on the average gradient over a subset of the training points, balancing computational efficiency and stability.

Note that the loss $\ell(f_\theta(Y^j), Z^j)$ is assumed to be differentiable with respect to θ. However, in our optimization problem, the ruin probability imposes numerical challenges due to the non-smooth nature of the indicator function. To address this issue, Subsection 4.1.1 explores connections to binary classification by replacing the indicator function with a surrogate loss function. This substitution makes the problem tractable via empirical risk minimization. Our surrogate loss model can be interpreted as a generalized (and to the best of our knowledge, never used before) version of a Gerber–Shiu penalty function.

4.1.1. Ruin probability and binary classification problems

Ruin probabilities can be expressed in terms of expectations. To this end, given $b \in \mathcal{A}$, consider the map F_b defined by $F_b(Y) = \min_{0 \le i \le n} X_i^b$. Then, one writes:

$$\mathbb{P}(\min_{0 \le i \le n} X_i^b < 0) = \mathbb{P}(F_b(Y) < 0) = \mathbb{E}[\mathbb{1}_{(-\infty,0)}(F_b(Y))], \tag{4.3}$$

where $\mathbb{1}_A(x)$ denotes the indicator function over the set A, i.e. $\mathbb{1}_A(x) = 1$ for $x \in A$ and $\mathbb{1}_A(x) = 0$ otherwise.

Minimizing the ruin probability in (4.3) then amounts to finding the optimal $b \in \mathcal{A}$, such that the mapping F_b classifies as many data points $Y(\omega)$ as possible not as ruin. In other words, finding the optimal reinsurance strategy $b \in \mathcal{A}$ can be identified with the equivalent task of finding the optimal classifier $F_{b^*} \in \{F_b \colon b \in \mathcal{A}\}$ for the binary classification problem where one seeks to map all data points $\{Y(\omega) \colon \omega \in \Omega\}$ to a non-ruin event. Using Y as an argument to F_b is justified by the Doob–Dynkin representation theorem, which states that all $\mathbb{F}$-adapted processes can be written as functions of Y.

The main issue with the indicator function is that it is neither convex nor smooth. Therefore, optimizing (4.3) by using deep learning tools and empirical risk minimization, particularly minibatch stochastic gradient descent, becomes quite challenging. A remedy to this issue is to employ a *surrogate loss function* g instead of the indicator function:

$$\mathbb{E}[\mathbb{1}_{(-\infty,0)}(F_b(Y))] \approx \mathbb{E}[g(F_b(Y))]. \tag{4.4}$$

The surrogate loss function $g_\gamma(x) = 0.5 + 0.5 \tanh(-\gamma x)$ for $\gamma \in \{1, 10, 100\}$ is presented in Figure 4.1. This function will be also used (with $\gamma = 10$) in the numerical study in

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Section 4.3. For theory on surrogate loss functions for binary classification problems, see Bartlett et al. (2006); Nguyen et al. (2009); Reid and Williamson (2010).

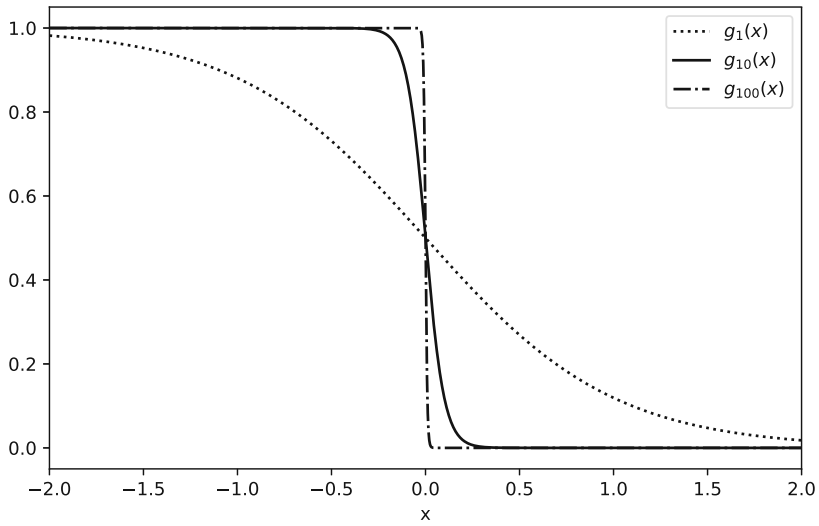

Figure 4.1. *Surrogate loss functions g_γ for various choices of γ.*

Replacing the ruin probability in (4.2) by the expected surrogate loss of $F_b(Y)$ with respect to g_γ imposes a penalty on the expected utility, which can be seen as a generalized Gerber–Shiu function (Gerber and Shiu, 1998). Indeed, our surrogate loss function allows for an approximation of the ruin probability under mild assumptions.

Proposition 4.4. *Given $b \in \mathcal{A}$, let $(g_n)_{n\in\mathbb{N}}$ be a uniformly bounded sequence of functions such that, $\mathbb{P}$-almost surely, $g_n(F_b(Y)) \to \mathbb{1}_{(-\infty,0)}(F_b(Y))$. Then,*

$$\lim_{n\to\infty} \mathbb{E}[g_n(F_b(Y))] = \mathbb{P}(\min_{0\le i\le n} X_i^b < 0).$$

Proof. This is a direct consequence of Lebesgue's dominated convergence theorem. □

Remark 4.5. For our choice g_γ of surrogate loss function as presented in Figure 4.1, Proposition 4.4 is applicable as $\gamma \to \infty$ if we assume that $F_b(Y)$ does not have a point mass in zero, i.e. we require $\mathbb{P}(F_b(Y) = 0) = 0$. The reason for that is that $g_\gamma(0) = 1/2$, which is the only point where g_γ does not converge to the indicator function $\mathbb{1}_{(-\infty,0)}$.

4.2. Algorithmic reinsurance policies

We propose to solve (4.2) numerically via *algorithmic reinsurance policies*. These policies determine retention levels using neural networks that observe information to generate

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

decisions. This approach is inspired by recent successful applications in quantitative finance and actuarial science. Popular use-cases are hedging, optimal stopping, model calibration, and scenario generation (Buehler et al., 2019; Becker et al., 2019; Horvath et al., 2021; Wiese et al., 2020).

Theorems that establish approximations in function space via neural networks are usually referred to as *universal approximation theorems* (UAT); notable contributions include Cybenko (1989) and Hornik (1991). These theorems establish the topological density of sets of neural networks in various topological spaces. One speaks of the universal approximation property (Kratsios, 2021) of a class of neural networks. Unfortunately, these theorems are usually non-constructive. To numerically find optimal neural networks, one typically combines backpropagation (see for example Rumerlhart et al. (1986)) with ideas from stochastic approximation (Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Dvoretzky, 1956).

Definition 4.6 (Deep feedforward neural network)**.** Let $\psi\colon \mathbb{R} \to \mathbb{R}$ be a bounded, measurable and non-constant map. Given $k, l, n \in \mathbb{N}$, we denote by $\mathcal{NN}_{k,l,n}(\psi)$ the set of neural networks with an n-dimensional input layer, one neuron with identity activation function in the output layer, k hidden layers, and at most l nodes with ψ as activation function in each hidden layer (cf. Kidger and Lyons (2020, Definition 3.1)). We call elements from $\mathcal{NN}_{k,l,n}(\psi)$ *deep feedforward neural networks*, or simply *deep neural networks*.

Definition 4.7 (Algorithmic reinsurance policy)**.** We denote by $\mathcal{A}^{\mathrm{nn}}$ the set of all proportional reinsurance strategies b that satisfy $b_i = \sigma \circ f(Y_0, Y_1, \ldots, Y_i)$ for some $f \in \mathcal{NN}_{k_i,l_i,(i+1)r}(\psi)$ with $k_i, l_i \in \mathbb{N}$, for each $i = 0, 1, \ldots, n$. Here, $\sigma\colon \mathbb{R} \to (0,1)$ denotes the logistic function given by $\sigma(x) = \exp(x)/(1+\exp(x))$.

Definition 4.6 provides one example of a class of neural networks we can use, but other choices are possible. In light of Definition 4.7, one could consider for example recurrent neural networks or long short-term memory (LSTM) networks (see for example Hochreiter and Schmidhuber (1997)). We also refer to UATs for deep, narrow networks (Kidger and Lyons, 2020) and for randomized neural networks (Huang et al., 2006). Note that feedforward neural networks are usually defined with a linear readout map (as in Definition 4.6). In order to ensure that algorithmic reinsurance policies are valid proportional reinsurance policies assuming values in $[0,1]$, we opt for the composition with the logistic function σ in Definition 4.7.

In this chapter, we restrict ourselves to proportional reinsurance strategies. However, the same techniques can be applied to other types of reinsurance strategies. For example, excess-of-loss (XL) policies could be written as $b_i(Z) = \min\{Z, f(Y_0, Y_1, \ldots, Y_i)\}$ for some deep neural network f. We leave the investigation and comparison of results for various algorithmic reinsurance treaties to future research.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

For the sake of notational simplicity, given $b \in \mathcal{A}$, $\beta \in [0,1]$ and a smooth surrogate loss function g, let

$$u_\beta(Y,b) = \beta u(X_n^b) - (1-\beta) g(F_b(Y)). \tag{4.5}$$

From a numerical perspective, solving the optimization problem (4.2) via algorithmic reinsurance policies motivates the use of Monte Carlo methods. Here, we first generate a finite amount of data points $(Y^j)_{j=1}^m$ and replace, for each $b \in \mathcal{A}$, the expectation and probability appearing in (4.2) by the empirical average

$$\frac{1}{m} \sum_{j=1}^{m} u_\beta(Y^j, b). \tag{4.6}$$

The summation over a finite number $m \in \mathbb{N}$ of data points allows us to re-interpret Eq. (4.6) as the expectation of $u_\beta(Y,b)$ over a measure which assigns equal probability $1/m$ to every outcome. This motivates the assumption that the underlying probability space Ω is finite, in which case all reinsurance strategies, including all algorithmic reinsurance policies, are admissible, i.e. in particular $\mathcal{A}^{\mathrm{nn}} \subset \mathcal{A}$. Moreover, since in this case every singleton set $\{\omega\}$ for each $\omega \in \Omega$ is measurable (for otherwise we could not assign the probability $1/m$ to it), it is also natural to assume that Ω is endowed with the power set $\mathcal{P}(\Omega)$ to form a measurable space.

Theorem 4.8. *Assume that Ω is a finite set, and that $\mathcal{F} = \mathcal{P}(\Omega)$. Then, for every $\beta \in [0,1]$ and $\varepsilon > 0$, there exists an algorithmic reinsurance policy $b^{\mathrm{nn}} \in \mathcal{A}^{\mathrm{nn}}$ such that*

$$\mathbb{E}[u_\beta(Y, b^{\mathrm{nn}})] > \sup_{b \in \mathcal{A}} \mathbb{E}[u_\beta(Y,b)] - \varepsilon. \tag{4.7}$$

Proof. The proof relies on some ideas from the proof of Buehler et al. (2019, Proposition 4.3). Let $b^* \in \mathcal{A}$ be an $\varepsilon/2$-optimal strategy, i.e. $\mathbb{E}[u_\beta(Y,b^*)] > \sup_{b\in\mathcal{A}} \mathbb{E}[u_\beta(Y,b)] - \varepsilon/2$. Fix one time point t_i. Since b^* is $\mathbb{F}$-adapted, we have that b_i^* is $\mathcal{F}_i$-measurable. Recall that $\mathcal{F}_i$ is the smallest σ-algebra that makes $Y_0, Y_1, \dots, Y_i$ measurable. An application of Doob–Dynkin's lemma implies the existence of a Borel-measurable map $f_i \colon \mathbb{R}^{r\times(i+1)} \to [0,1]$ such that $b_i^* = f_i(Y_0, Y_1, \dots, Y_i)$.

Let μ be the Borel probability measure that is induced by $Y_0, Y_1, \dots, Y_i$ on $\mathbb{R}^{r\times(i+1)}$. Since b_i^* is bounded (as it assumes values in $[0,1]$), we have $b_i^* \in L^p(\mathbb{P})$ for every $p > 0$ and thus, in particular, $f_i \in L^2(\mu)$.

Consider the sequence $(f_i^k)_{k\in\mathbb{N}}$ of functions given by $f_i^k(x) = 1 - 1/k$ if $f_i(x) = 1$, $f_i^k(x) = 1/k$ if $f_i(x) = 0$, and $f_i^k(x) = f_i(x)$ otherwise. Clearly, $f_i^k \to f_i$ pointwise and thus, $f_i^k(Y_0, Y_1, \dots, Y_i) \to f_i(Y_0, Y_1, \dots, Y_i) = b_i^*$ pointwise. We repeat the same construction for all time points, and construct a sequence $(\hat{b}^k)_{k\in\mathbb{N}}$ of reinsurance policies, where $\hat{b}_i^k = f_i^k(Y_0, Y_1, \dots, Y_i)$ for each $i = 0, 1, \dots, n-1$. Since Ω is finite,

$$\mathbb{E}[u_\beta(Y, \hat{b}^k)] \to \mathbb{E}[u_\beta(Y, b^*)], \quad k \to \infty.$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

We may therefore assume, without loss of generality, that f_i takes values in a closed subset $[1/k, 1-1/k]$ of $[0,1]$ that does neither contain 0 nor 1, for a sufficiently large $k \in \mathbb{N}$.

The sigmoid function σ, being a continuous bijection with continuous inverse $\sigma^{-1}\colon (0,1) \to \mathbb{R}$, establishes a homeomorphism between $\mathbb{R}$ and $(0,1)$. Since $f_i \in [1/k, 1-1/k]$, it follows that $\sigma^{-1} \circ f_i$ is bounded, and thus $\sigma^{-1} \circ f_i \in L^2(\mu)$.

The universal approximation theorem (Hornik, 1991) ensures the existence of a sequence $(\tilde{f}_i^k)_{k\in\mathbb{N}}$ of shallow feedforward neural networks (i.e. feedforward neural networks with one hidden layer), such that $\tilde{f}_i^k \to \sigma^{-1} \circ f_i$ in $L^2(\mu)$, as $k \to \infty$. But this implies, denoting $\tilde{b}_i^k = \tilde{f}_i^k(Y_0, Y_1, \dots, Y_i)$, that $\tilde{b}_i^k \to \sigma^{-1} \circ b_i^*$ in $L^p(\mathbb{P})$, as $k \to \infty$, and thus also in probability. Upon passing to a subsequence, which we also denote $(\tilde{b}_i^k)_{k\in\mathbb{N}}$, we may assume that convergence holds outside of a $\mathbb{P}$-null set, in which case $\sigma \circ \tilde{b}_i^k \to b_i^*$, $\mathbb{P}$-almost surely.

If we repeat the above arguments for every t_i, we obtain a sequence of processes $(b^k)_{k\in\mathbb{N}}$ given by $b_i^k = \sigma \circ \tilde{b}_i^k$, such that $b_i^k \to b_i^*$ for every $i = 0, 1, \dots, n$, where convergence holds outside a $\mathbb{P}$-null set.

This implies that $X_i^{b^k} \to X_i^{b^*}$ for every $i = 0, 1, \dots, n$, where convergence holds outside a $\mathbb{P}$-null set, and consequently $F_{b^k}(Y) \to F_{b^*}(Y)$, $\mathbb{P}$-almost surely, which implies that $u(X_n^{b^k}) \to u(X_n^{b^*})$ and $g(F_{b^k}(Y)) \to g(F_{b^*}(Y))$, $\mathbb{P}$-almost surely as $k \to \infty$. Since Ω is finite,

$$\mathbb{E}[u_\beta(Y, b^k)] \to \mathbb{E}[u_\beta(Y, b^*)], \quad k \to \infty.$$

We can thus find k^* large enough, such that $|\mathbb{E}[u_\beta(Y, b^*)] - \mathbb{E}[u_\beta(Y, b^{k^*})]| < \varepsilon/2$, and set $b^{\mathrm{nn}} = b^{k^*}$. □

Remark 4.9. Theorem 4.8 provides theoretical justification for solving the optimization problem (4.2) via algorithmic reinsurance policies. However, it is non-constructive in the sense that it does not shed any light on the way how the policy b^{nn} can be found. In order to solve the problem via empirical risk minimization in the next section, we will proceed in two steps:

(a) Approximate the indicator function used to form the ruin probability by a surrogate loss function. This step is justified by Proposition 4.4.

(b) Solve the surrogate problem by stochastic approximation over the set $\mathcal{A}^{\mathrm{nn}}$ of algorithmic policies. This step is justified by Theorem 4.8.

The final result is subject to two sources of numerical error: (1) the approximation of the indicator function, and (2) the approximation of the solution to the surrogate problem with algorithmic policies. However, both of these two errors can be made arbitrarily small.

Remark 4.10. Classical universal approximation theorems which are formulated for feedforward neural networks usually assume a linear readout, which is not bounded. In

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

order to obtain a valid proportional reinsurance policy, we need to guarantee that our algorithmic strategies assume values in $[0, 1]$. In Theorem 4.8 we were able to do this with some simple tricks for our specific model setting. For a more general treatment of non-Euclidean universal approximation, see Kratsios and Bilokopytov (2020).

4.3. Numerical study

In this section, we present a numerical example where the surplus is modeled by a Cramér–Lundberg model perturbed by an Ornstein–Uhlenbeck (OU) process. Including the OU process adds complexity, making the problem more challenging and realistic. The OU process may, for instance, represent the fluctuations in the number of clients or in premium payments.

We assume that the claims Z are iid, exponentially distributed with mean μ, N represents a discretized Poisson process with intensity λ, and Z and N are independent. The insurer charges premia according to the expected value principle with safety loading $\eta > 0$, i.e. $p_i = (1 + \eta)\lambda\mu(t_{i+1} - t_i)$. Similarly, the reinsurer charges premia according to the expected value principle with safety loading $\theta > \eta$, i.e. $c_i = (1+\theta)\lambda\mu(1-b_i)(t_{i+1}-t_i)$. The process L follows the dynamics

$$L_{i+1} = L_i + \xi(\kappa - L_i)(t_{i+1} - t_i) + \nu(t_{i+1} - t_i)\varepsilon_i, \tag{4.8}$$

where ε_i are iid shocks with $\mathcal{L}(\varepsilon_i) = \mathcal{N}(0, 1)$. We assume that the initial capital is positive to avoid starting in ruin. For our simulations, we take an arbitrarily chosen initial value of 1.

The utility function is of exponential type, $u(x) = \exp(-\alpha x)$, with risk-aversion coefficient α. Table 4.1 summarizes the parameters for our base model. The values for λ, μ, η and θ are taken from Schmidli (2001). For the neural network, we chose a two-hidden-layer topology with hyperbolic tangent (tanh) as activation function in the hidden layers and logistic activation σ in the output layer. The hidden layers contain 32 nodes each. The neural network takes the surplus as input; additional inputs were tested but provided negligible improvements. All computations were performed using Python, using the Keras deep learning API for constructing and training the neural networks.

Neural network training was performed on 2000 batches, each with a batch size of 2^{14}, using the Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of 10^{-3}. The learning rate was decreased by a factor of 10 after 10 epochs without improvement, with a minimum learning rate of 10^{-5}. Early stopping was employed after 20 epochs without improvement. Distributions, expected utilities, and ruin probabilities were computed on a test set of size 2^{25}.

First, we want to numerically verify that the expected surrogate loss indeed approximates the ruin probability. To achieve this, we compute the ruin probability for our model without reinsurance (i.e. setting $b \equiv 1$) and compare the obtained value ($\approx 34.1\%$) with

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 4.1. *Parameters for the base model.*

Model parameter	**Value**
Initial wealth	$x = 1$
Time horizon	$T = 10$
Number of time steps	$n = 10$
Claim arrival intensity	$\lambda = 1.0$
Expected claim size	$\mu = 1.0$
Safety loading of insurer	$\eta = 0.50$
Safety loading of reinsurer	$\theta = 0.70$
Risk aversion coefficient	$\alpha = 0.30$
Tuning parameter	$\beta = 0.40$
Surrogate loss parameter	$\gamma = 10$
OU mean-reversion level	$\kappa = 0$
OU mean-reversion speed	$\xi = 0.20$
OU volatility coefficient	$\nu = 0.05$

the expected surrogate loss $\mathbb{E}[g_\gamma(F_b(Y))]$ for various values of γ. As shown in Figure 4.2, the expected surrogate loss approximates the ruin probability well when γ is sufficiently large. Based on these results, we fix $\gamma = 10$ for the subsequent analysis.

The parameter β, given in (4.2), significantly influences the optimal retention level as a function of the surplus. For instance, in Schmidli (2001), where the ruin probability is minimized in a continuous-time setting, the optimal reinsurance strategy turns out to be constantly 1 until a certain surplus level is reached. After that, the retention level drops to a much lower value and converges to a constant as the surplus goes to infinity. In our model, by incorporating the ruin probability as an additional objective, the reinsurance policy must penalize scenarios where the insurer's wealth becomes negative at any point in time up to maturity T.

In our simulations, we noticed that it sufficed to optimize over those feedback strategies $b = (b_i)_{i=0}^n$ that do not depend on time, i.e. $b_i(x) = \bar{b}(x)$ for $i = 0, 1, \ldots, n-1$ (with $b_n \equiv 1$) and for some function $\bar{b}\colon \mathbb{R} \to [0, 1]$, as using time-dependent strategies did not improve the solution. Therefore, we can interpret our obtained strategies as functions depending on the initial capital. Figure 4.3 shows the optimal retention levels depending on the initial capital for the base case $\beta = 0.4$ and the boundary cases $\beta = 1$ (no penalisation through ruin probability, only utility maximization) and $\beta = 0$ (pure ruin probability minimization, no utility in the target functional).

One can check that the strategy corresponding to the case of pure ruin probability minimization has the same form (on the positive real line) as the strategy in the continuous time setting for the classical risk model, see Schmidli (2008, p. 53). In the base model, the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

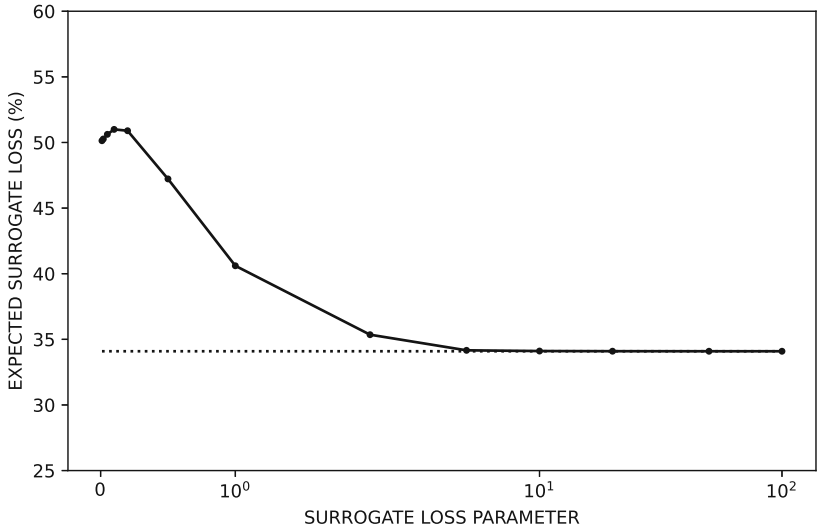

Figure 4.2. *Expected surrogate loss* $\mathbb{E}[g_\gamma(F_b(Y))]$ *for various parametrizations of the surrogate loss function in the case of no reinsurance* ($b \equiv 1$). *The dotted horizontal line corresponds to the ruin probability* ($\approx 34.1\%$).

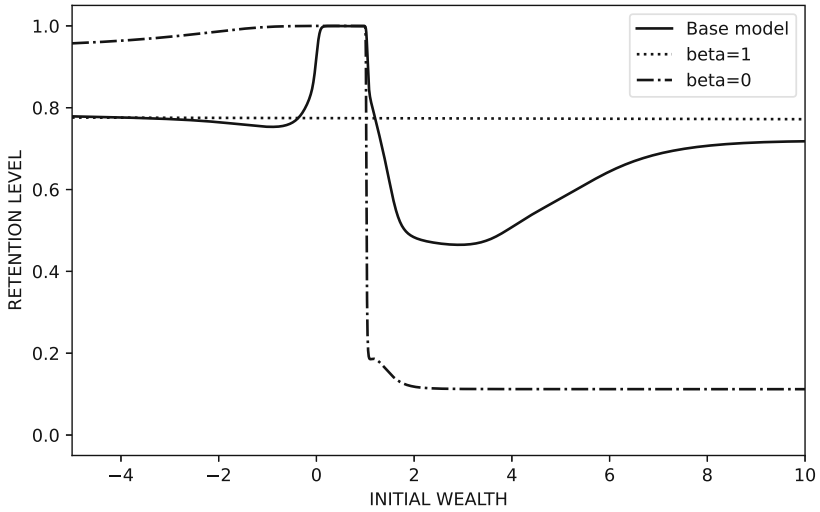

Figure 4.3. *Optimal retention levels* $\bar{b}(x)$ *for different values of* β. *The base model assumes* $\beta = 0.4$.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

presence of the utility function in the target functional enforces an increase in the optimal retention level starting from approximately $x = 3$. In the tradeoff between more safety (a smaller retention level) and more utility, the utility wins in the long run. Obviously, by maximizing exponential utility with no consideration of the ruin probability, the optimal strategy does not depend on the surplus (dotted line in Figure 4.3).

Recall that our objective contains two components: the expected utility of terminal wealth, and the expected surrogate loss approximating the ruin probability. By varying β, we can generate optimal reinsurance strategies that approximately achieve Pareto-efficient solutions, where any improvement in one objective requires a compromise in the other. It is important to note that our results are subject to numerical errors due to:

(a) The surrogate loss function for the ruin probability,

(b) The finite size of training and test datasets, and

(c) The fact that neural networks can in general only approximate optimal policies up to some ε.

Figure 4.4 demonstrates the trade-off between expected utility of terminal wealth and ruin probability. Each individual point represents an approximate Pareto-efficient solution for different choices of $\beta \in [0, 1]$, where improving one objective comes at the expense of the other. The star indicates the values obtained without reinsurance (i.e. $b \equiv 1$), highlighting the benefits of optimal reinsurance strategies. In particular, the figure highlights that, for our choice of parameters, reinsurance is always more favourable than non-reinsurance.

Conclusions

We introduce a novel framework for optimizing reinsurance strategies using a deep learning approach. The target functional consists of the expected utility of terminal wealth perturbed by a modified Gerber–Shiu penalty function. It allows to balance between maximizing the expected utility of terminal wealth, and minimizing the probability of ruin, depending on the individual preferences of the insurer.

We draw connections to concepts from binary classification and surrogate loss functions. This enables the problem to be addressed using empirical risk minimization methods. Combined with stochastic gradient descent, it allows for efficient optimization of algorithmic reinsurance policies, even in complex model settings.

Our numerical findings highlight the ability of our method to interpolate between the problems of maximizing expected utility of terminal wealth, and minimizing the probability of ruin. Future research could explore other optimization targets and reinsurance forms as well as more complex, higher-dimensional models with correlated business lines. Moreover, it would be interesting to explore optimal algorithmic reinsurance strategies

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

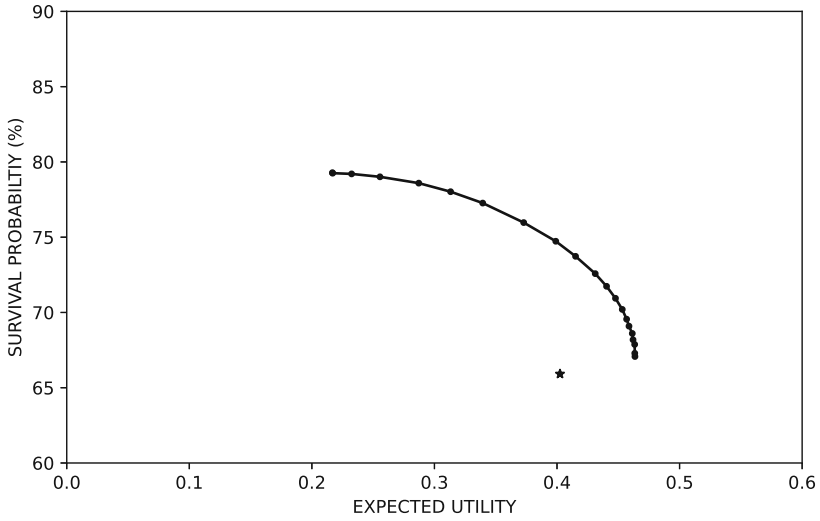

Figure 4.4. *Trade-off between expected utility of terminal wealth and survival probability obtained through optimal algorithmic reinsurance polices for different choices of* $\beta \in [0, 1]$. *The star denotes the corresponding values obtained when no reinsurance is available.*

that are robust with respect to parameter uncertainties, such as the intensity of the Poisson process, or the expected claim size. Finally, another interesting topic would be to optimize algorithmic reinsurance polices under distributional constraints on the wealth process.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

5. Algorithmic Strategies in Continuous-Time Hedging and Stochastic Integration

The starting point for this chapter is Buehler et al. (2019). In that work, the authors present a tractable framework for constructing ε-optimal hedging strategies via machine learning technology. One of the key innovations is a simple yet profound application of Doob's functional representation lemma which allows to represent trading strategies via neural networks. The framework is formulated in great generality, allowing for the consideration of transaction costs and liquidity constraints. However, by choosing a discrete-time rather than a continuous-time setting, the authors circumvent many technicalities such as semimartingale properties of the asset price process, predictability of the trading strategies, and necessary properties of the filtration which carries all market information.

Around the late 80s and early 90s, it has been shown that neural networks are a class of universal approximators. Some notable contributions include Cybenko (1989), Funahashi (1989), Hornik et al. (1989), Hecht-Nielsen (1989), Hornik (1991) and Leshno et al. (1993), see Schmidhuber (2015) for an extensive survey. What these works showed in particular, is that neural networks constitute topologically dense subsets of $C(K)$ and $L^p(\mu)$, the spaces of real-valued continuous functions on a compact domain K, and real-valued $L^p(\mu)$-integrable functions with respect to a finite Borel measure μ, respectively. This functional analytic insight means that every function in $C(K)$ and $L^p(\mu)$ can – in theory – be approximated arbitrarily well by a neural network with respect to the respective notion of distance on these spaces, namely uniform- and $L^p(\mu)$-distance. Results like these are commonly referred to as *universal approximation theorems*, and one also speaks of the *universal approximation property* of neural networks (Kratsios, 2021).

The spaces $C(K)$ and $L^p(\mu)$ have their natural counterparts in the theory of stochastic calculus. Here, one speaks of ucp- or S^p-convergence of stochastic processes, depending whether the uniform distance – either uniformly over all time points in case the index set is bounded, or uniformly over any arbitrary compact time interval in case the index set is unbounded – to a given process converges to zero, in probability or in a given L^p-space, respectively. Usually, one restricts to a subclass of stochastic processes that carry special properties, like being semimartingales, or having paths with certain regularity properties. We refer to the excellent textbooks by He et al. (1992), Protter (2005), Cohen and Elliott

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

(2015) and Kallenberg (2021) for precise definitions and many results centred around with ucp- and S^p-convergence.

At this point it is only natural to ask whether the universal approximation theorems and approximations of optimal trading and hedging strategies as discussed in Buehler et al. (2019) can be transferred to spaces of stochastic processes in continuous time, and this is precisely our goal. To this end, we consider algorithmic strategies, which are linear combinations of algorithmic buy-and-hold strategies,

$$f(\tau_i, Y_{t_1\wedge\tau_i}, \dots, Y_{t_n\wedge\tau_i})\mathbb{1}_{(\tau_i,\tau_{i+1}]}.$$

At (random) times τ_i, a neural network f observes finitely many realizations of the available and past market information, $Y_{t_1\wedge\tau_i}, \dots, Y_{t_n\wedge\tau_i}$, and enters into a buy-and-hold trade over the amount $f(\tau_i, Y_{t_1\wedge\tau_i}, \dots, Y_{t_n\wedge\tau_i})$ that is executed until some later (random) time $\tau_{i+1} > \tau_i$. In that way, algorithmic strategies form subset of so called simple predictable step processes, which play a fundamental role in the construction of stochastic integrals.

In the same way that measure theory uses step functions to define an elementary integral, and then extend this notion to the integral of more general functions by passing to a limit, stochastic integrals of integrable processes are commonly defined as limits over sequences of stochastic integrals of simple processes. By proving that a large class of stochastic processes – including simple processes – can be approximated by algorithmic strategies in terms of ucp- and S^p-convergence, we obtain as a consequence that stochastic integral processes can be approximated by integrals of algorithmic strategies too (see Section 5.3 for a proof):

Theorem 5.1. *Let X be a càdlàg semimartingale, and V a càglàd process. Then, there exists a sequence $(V^n)_{n\in\mathbb{N}}$ of algorithmic strategies that converges to V, and such that*

$$\int V^n \,\mathrm{d}X \to \int V \,\mathrm{d}X, \quad n \to \infty,$$

where convergence in both cases holds in the topology of ucp-convergence.

Theorem 5.1 highlights the fundamental role that neural networks – through algorithmic strategies – play for the theory of stochastic integration, extends the setting of Buehler et al. (2019) to the continuous-time domain, and yields several interesting implications from the point of view of mathematical finance.

Integrals of algorithmic strategies with respect to square-integrable martingale price processes are topologically dense in the stable subspace generated by these price processes. As such, algorithmic strategies can replicate the performance of mean-variance optimal hedging strategies (Kunita and Watanabe, 1967; Ansel and Stricker, 1993; Rheinländer and Schweizer, 1997) arbitrarily well, giving rise to the notion of deep mean-variance

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

hedging in continuous time, and extending the corresponding result of Buehler et al. (2019) from the discrete- to the continuous-time domain (see Theorem 5.35 below).

Stricker (1990); Delbaen (1992); Delbaen and Schachermayer (1994) have linked the existence of an equivalent martingale measure for a price process to the question whether trading with simple processes allows for a free lunch. By refining Stricker (1990, Theorem 4), we show that a sufficient condition for the existence of an equivalent martingale measure for the price process can be linked to the absence of a free lunch with vanishing risk for algorithmic strategies (see Theorem 5.37 below).

To summarize, the primary contributions of this chapter are as follows:

(a) Theorem 5.16 extends the approximation capabilities of neural networks (Hornik, 1991, Theorem 1) from L^p-spaces to the more general class of Orlicz spaces of real-valued functions over domains that are general locally convex spaces.

(b) Theorem 5.23 demonstrates that, in case the underlying σ-algebra is generated by an (uncountable) family of random variables, neural networks – through functional representations – can approximate functions in these Orlicz spaces arbitrarily well.

(c) By representing algorithmic strategies as simple predictable processes, Theorems 5.29 and 5.32 establish their approximation capabilities in spaces of stochastic (integral) processes.

(d) Theorem 5.35 extends the notion of deep mean-variance hedging to the continuous-time setting, by demonstrating that algorithmic strategies can approximate mean-variance optimal hedging strategies arbitrarily well.

(e) Finally, Theorem 5.37 establishes a no free lunch with vanishing risk condition for algorithmic strategies.

The structure of the chapter is as follows: Section 5.1 discusses universal approximation in Orlicz spaces. Section 5.2 introduces algorithmically generated random variables. Section 5.3 introduces the corresponding algorithmic strategies and discusses their approximation capabilities.

5.1. Neural networks on locally convex spaces

One of the most prominent features of neural networks is their universal approximation property. In Hornik (1991, Theorem 1), a universal approximation theorem in $L^p(\mu)$-spaces of real-valued functions with $p \in [1, \infty)$ for finite Borel measures μ on the Euclidean space $\mathbb{R}^d$ with $d \in \mathbb{N}$ was proved. In this section, we extend Hornik (1991, Theorem 1) to Orlicz spaces with respect to finite measures on locally convex Hausdorff topological vector spaces (for brevity, locally convex spaces), see Theorem 5.16 below.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In order to introduce the concept of a neural network, let $(X, \mathcal{A})$ be a measurable space, and Γ a non-empty set of real-valued $\mathcal{A}$-measurable functions. For example, X might be a complete separable metric space with Borel σ-algebra $\mathcal{A} = \mathcal{B}_X$, and Γ could be a set of real-valued $\mathcal{B}_X$-measurable functions that is point-separating for X. In that case, Bogachev (2007, Theorem 6.8.9) ensures that Γ generates $\mathcal{B}_X$, i.e. $\mathcal{B}_X = \sigma(f\colon f \in \Gamma)$. In locally convex Hausdorff topological vector spaces, there always exists a point-separating set Γ, as the Hahn–Banach theorem (Rudin, 1991, Theorem 3.4) shows that X^*, the continuous dual space, is point-separating for X. However, for the time being, we do not restrict the measurable space $(X, \mathcal{A})$, only later will we impose Standing Assumption 5.13(a) for purpose of proving Theorem 5.16.

Definition 5.2 (Neural network)**.** Let $\psi\colon \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function. We define $\mathcal{NN}(\Gamma, \psi)$ as the $\mathbb{R}$-linear span of functions of the form

$$x \mapsto \psi(f(x) + \beta), \tag{5.1}$$

where $f \in \Gamma$, and $\beta \in \mathbb{R}$. The elements of $\mathcal{NN}(\Gamma, \psi)$ are referred to as (shallow feedforward) neural networks with activation function ψ.

Neural networks according to Definition 5.2 are real-valued $\mathcal{A}$-measurable functions on X. We briefly describe the extension to neural networks with values in an infinite-dimensional Banach space. Let Y be a real Banach space with a Schauder basis $(e_i)_{i\in\mathbb{N}}$, meaning that for each $y \in Y$, there exists a unique sequence $(a_i)_{i\in\mathbb{N}}$ of real coefficients, such that $y = \sum_{i=1}^{\infty} a_i e_i$, where the series converges in Y with respect to the topology induced by the norm on Y. The set $\mathcal{NN}(\Gamma, \psi; Y)$ of Y-valued neural networks on X is then defined as the $\mathbb{R}$-linear span of functions of the form

$$x \mapsto \sum_{i\in I} \psi(f_i(x) + \beta_i) e_i,$$

where $I \subset \mathbb{N}$ is a finite index set, $f_i \in \Gamma$ and $\beta_i \in \mathbb{R}$ for all $i \in I$. By construction, elements of $\mathcal{NN}(\Gamma, \psi; Y)$ assume values in finite-dimensional subspaces of Y. This definition is also consistent with the case where $Y = \mathbb{R}^n$ with $n \in \mathbb{N}$, in which case the Schauder basis $(e_i)_{i\in\mathbb{N}}$ is replaced by a basis $\{e_1, \dots, e_n\}$ of $\mathbb{R}^n$, and the index set I is constrained to the finite set $\{1, \dots, n\}$.

Remark 5.3. Let $g \in \mathcal{NN}(\Gamma, \psi)$. Then, it admits a representation of the form

$$g(x) = \sum_{i=1}^{n} \alpha_i \psi(f_i(x) + \beta_i), \tag{5.2}$$

where $n \in \mathbb{N}$ is the number of hidden nodes, $\alpha_1, \dots, \alpha_n$ are real weights, $\beta_1, \dots, \beta_n$ are real biases, and $f_1, \dots, f_n \in \Gamma$. For the moment, let us assume that the activation function ψ is bounded, that X is a real separable Banach space, and that $\Gamma = X^*$. Let μ

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

be a finite signed Borel measure on $\Theta \coloneqq \Gamma \times \mathbb{R}$. If we assume Γ to be separable (when endowed with the topology induced by the operator norm), then Θ equipped with the product topology is separable, too, and the product σ-algebra $\mathcal{B}_\Gamma \otimes \mathcal{B}_\mathbb{R}$ is equal to $\mathcal{B}_\Theta$ (Schmock, 2024, Theorem 15.54). We then extend g in Eq. (5.2) to an infinite-width neural network

$$g(x) = \int_\Theta \psi(f(x) + \beta)\, \mu(\mathrm{d}f, \mathrm{d}\beta), \tag{5.3}$$

and we recover the representation (5.2) by choosing $\mu = \sum_{i=1}^n \alpha_i \delta_{z_i}$, with $z_i = (f_i, \beta_i)$.

Remarkably, fitting infinite-width neural networks to data via an infinite-dimensional convex optimization problem subject to a penalty on the total-variation norm of μ can result in sparse minimizers of the form $\mu = \sum_{i=1}^n \alpha_i \delta_{z_i}$, with $n \in \mathbb{N}$ and $z_i = (f_i, \beta_i) \in \Theta$ for $i = 1, \dots, n$ (Bach, 2017). Moreover, the integral representation (5.3) can further be related to the Radon transform (Unser, 2023).

Remark 5.4. In the context of Remark 5.3, it is not always the case that Γ is separable; there are non-trivial examples of real separable Banach spaces whose topological dual is not separable. For instance, consider the probability space $([0,1], \mathcal{B}_{[0,1]}, \lambda)$, where λ denotes the Lebesgue–Borel measure on $[0,1]$. We know that $(L^1(\lambda))^* \simeq L^\infty(\lambda)$ (Billingsley, 2012, Theorem 19.3), that both $L^1(\lambda)$ and $L^\infty(\lambda)$ are Banach spaces (Billingsley, 2012, Theorem 19.1), and – since $\mathcal{B}_{[0,1]}$ is countably generated – that $L^1(\lambda)$ is separable (Billingsley, 2012, Theorem 19.2). However, $L^\infty(\lambda)$ is not separable. For $\varepsilon \in (0,1]$, let $f_\varepsilon = 2\mathbb{1}_{[0,\varepsilon]}$. The set $\{f_\varepsilon \colon \varepsilon \in (0,1]\}$ contains uncountably many bounded, $\mathcal{B}_{[0,1]}$-measurable functions, and $\|f_\varepsilon - f_{\tilde{\varepsilon}}\|_{L^\infty(\lambda)} = 2$ for all $\varepsilon, \tilde{\varepsilon} \in (0,1]$ with $\varepsilon \neq \tilde{\varepsilon}$. This implies that the open balls $(B_1(f_\varepsilon))_{\varepsilon \in (0,1]}$ of unit radius are pairwise disjoint, hence there cannot exist a countable dense subset of $L^\infty(\lambda)$.

Besides interpreting Eq. (5.2) as a discretization of the integral representation in Eq. (5.3), it has also become popular in the literature to consider the limit as $n \to \infty$ in Eq. (5.2), where the coefficients α_i, f_i and β_i are randomly sampled in such a way that a central limit theorem is applicable. In some cases, one obtains in the limit that $(g(x))_{x \in X}$ becomes a Gaussian process, while in other cases more general distributions are possible. We refer the interested reader to Neal (1996) for the Gaussian case, and to Der and Lee (2005) for the general case.

We now discuss several examples for X, Γ and the corresponding sets $\mathcal{NN}(\Gamma, \psi)$ of neural networks. In particular, Example 5.8 demonstrates how one can define neural networks on spaces of stochastic processes.

Example 5.5. Let K denote a compact Hausdorff topological vector space, and $X = C(K)$, the Banach space of real-valued continuous functions on K, endowed with the supremum norm $\|f\| = \sup_{x \in K} |f(x)|$. Let $\mathcal{M}_\mathrm{r}(K)$ denote the set of all signed Radon measures on K. The Riesz representation theorem shows that X^* can be identified with $\mathcal{M}_\mathrm{r}(K)$ (Bogachev, 2007, Theorem 7.10.4). More precisely, for each $F \in X^*$, there exists a unique $\mu \in \mathcal{M}_\mathrm{r}(K)$, such that $F(f) = \int_K f(x)\, \mu(\mathrm{d}x)$ for each $f \in X$. Let

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$\Gamma = X^* \simeq \mathcal{M}_{\mathrm{r}}(K)$. The set $\mathcal{NN}(\Gamma, \psi)$ of neural networks on $C(K)$ is then defined as the $\mathbb{R}$-linear span of functions of the form

$$f \mapsto \psi\Big(\int_K f(x)\, \mu(\mathrm{d}x) + \beta \Big),$$

where $\mu \in \mathcal{M}_{\mathrm{r}}(K)$ and $\beta \in \mathbb{R}$.

Example 5.6. Let $X = C(\mathbb{R}_+)$, the space of real-valued continuous functions on $\mathbb{R}_+$. For a null sequence $(c_n)_{n\in\mathbb{N}}$ in $(0,\infty)$, let

$$\varrho(f,g) = \max_{n\in\mathbb{N}} \big(c_n \wedge \sup_{t\in[0,n]} |f(t) - g(t)|\big), \quad f, g \in X.$$

Then ϱ is a metric for the topology of uniform convergence on compact subsets of $\mathbb{R}_+$, and (X, ϱ) is a Polish space (Schmock, 2024, Exercise 2.121). The coordinate projections π_t given by $\pi_t(f) = f(t)$ for $t \in \mathbb{R}_+$ are point-separating and generate $\mathcal{B}_X$. Let $\Gamma = \{\pi_t \colon t \in \mathbb{R}_+\}$. The set $\mathcal{NN}(\Gamma, \psi)$ of neural networks on $C(\mathbb{R}_+)$ is then defined as the $\mathbb{R}$-linear span of functions of the form

$$f \mapsto \psi(f(t) + \beta),$$

where $t \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$.

Example 5.7. Let $T > 0$ be a fixed time horizon, and $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space endowed with filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$, such that $\mathcal{F}_0$ contains all $\mathbb{P}$-null sets of $\mathcal{F}$. Let $M = (M_t)_{t\in[0,T]}$ be a continuous $\mathbb{R}^d$-valued local $(\mathbb{F}, \mathbb{P})$-martingale with deterministic covariation process $[M]$. Then there exists a $\mathcal{B}_{[0,T]}$-measurable function $\pi \colon [0,T] \to \mathbb{R}^{d\times d}$ with values in the positive semi-definite matrices of $\mathbb{R}^{d\times d}$, and a finite Lebesgue–Stieltjes measure μ on $\mathcal{B}_{[0,T]}$, such that $[M]_t = \int_0^t \pi(s)\, \mu(\mathrm{d}s)$ for $t \in [0,T]$ (Jacod, 1979, Section IV.2). Let Λ^2 denote the set of all equivalence classes of $\mathcal{B}_{[0,T]}$-measurable functions $f \colon [0,T] \to \mathbb{R}^d$ with $\int_0^T f^\top(s)\pi(s)f(s)\, \mu(\mathrm{d}s) < \infty$, where we identify two such functions f and g if $(f-g)^\top \pi(\cdot)(f-g) = 0$ up to a set of μ-measure zero.

Endowed with the inner product $\langle f, g\rangle = \int_0^T f(s)\pi(s)g(s)\, \mu(\mathrm{d}s)$, Λ^2 is a real separable Hilbert space that does not depend on the specific choice of (π, μ) that satisfy the integral representation of the covariation process (Arandjelović et al., 2025, Lemma 2.11). By the Fréchet–Riesz representation theorem (Schmock, 2024, Theorem 14.16), the continuous dual $(\Lambda^2)^*$ is isometrically isomorphic to Λ^2. More precisely, to each $F \in (\Lambda^2)^*$, there corresponds a unique function $g \in \Lambda^2$, such that $F(f) = \int_0^T g^\top(s)\pi(s)f(s)\, \mu(\mathrm{d}s)$ for each $f \in \Lambda^2$. Let $\Gamma = (\Lambda^2)^* \simeq \Lambda^2$. The set $\mathcal{NN}(\Gamma, \psi)$ of neural networks on Λ^2 is then defined as the $\mathbb{R}$-linear span of functions of the form

$$f \mapsto \psi\Big(\int_0^T g^\top(s)\pi(s)f(s)\, \mu(\mathrm{d}s) + \beta \Big),$$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

where $g \in \Lambda^2$ and $\beta \in \mathbb{R}$.

Example 5.8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space endowed with a right-continuous filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$, such that $\mathcal{F}_0$ contains all $\mathbb{P}$-null sets of $\mathcal{F}$, and let $\mathcal{F}_\infty = \bigvee_{t\in\mathbb{R}_+} \mathcal{F}_t$. In that case, every $(\mathbb{F}, \mathbb{P})$-martingale has a càdàg modification (Cohen and Elliott, 2015, Corollary 5.1.9), and we shall always implicitly consider such a version when speaking of an $(\mathbb{F}, \mathbb{P})$-martingale. Moreover, let us identify those processes which are indistinguishable. For a càdlàg process $M = (M_t)_{t\in\mathbb{R}_+}$ and $t \in [0, \infty]$, let $M_t^* \coloneqq \sup_{s\in[0,t]\cap\mathbb{R}_+} |M_s|$.

Given $p \in [1, \infty)$, let $X = \mathcal{H}^p$, the Banach space of all real-valued $(\mathbb{F}, \mathbb{P})$-martingales M with $\|M\|_{\mathcal{H}^p} \coloneqq \|M_\infty^*\|_p < \infty$ (Cohen and Elliott, 2015, Lemma 10.1.5). If $p \in (1, \infty)$, then $\|M\|_{\mathcal{H}^p}$ is equivalent to $\|M_\infty\|_p$ on $\mathcal{H}^p$ (Cohen and Elliott, 2015, Lemma 10.1.3), where M_∞ denotes the a.s. existing limit of M. Therefore, $\mathcal{H}^p$ can be identified with the Banach space $L^p(\mathbb{P})$ by the map which associates $M \in \mathcal{H}^p$ with its $\mathcal{F}_\infty$-measurable limit $M_\infty \in L^p(\mathbb{P})$. Moreover, if $\mathcal{F}_\infty$ is countably generated, then $L^p(\mathbb{P})$ and thus $\mathcal{H}^p$ are separable (Billingsley, 2012, Theorem 19.2).

Fix $q = p/(p-1)$. Since $(L^p(\mathbb{P}))^*$ is isometrically isomorphic to $L^q(\mathbb{P})$ (Billingsley, 2012, Theorem 19.3), it follows that $(\mathcal{H}^p)^*$ can be identified with $\mathcal{H}^q$. More precisely, to each $F \in (\mathcal{H}^p)^*$ there corresponds a unique $N \in \mathcal{H}^q$, such that $F(M) = \mathbb{E}[M_\infty N_\infty]$ for every $M \in \mathcal{H}^p$. Let $\Gamma = (\mathcal{H}^p)^* \simeq \mathcal{H}^q$. The set $\mathcal{NN}(\Gamma, \psi)$ of neural networks on $\mathcal{H}^p$ is then defined as the $\mathbb{R}$-linear span of functions of the form

$$M \mapsto \psi(\mathbb{E}[M_\infty N_\infty] + \beta),$$

where $N \in \mathcal{H}^q$ and $\beta \in \mathbb{R}$.

Note that the case $p = 1$ is not covered by this construction. This is because $(\mathcal{H}^1)^*$ cannot be identified with $\mathcal{H}^\infty$ (the space of real-valued $(\mathbb{F}, \mathbb{P})$-martingales that are almost surely uniformly bounded), but rather with the space $\mathcal{H}^{\mathrm{BMO}}$ of BMO-martingales. We refer the interested reader to Section 2 and Theorem 2.6 in Kazamaki (1994) for further details about BMO-martingales and the property $(\mathcal{H}^1)^* \simeq \mathcal{H}^{\mathrm{BMO}}$.

We have now introduced neural networks as functions on general measurable spaces. As a next step, we aim to investigate the universal approximation property of neural networks in the Orlicz space $L^\Phi(\mu)$ and Orlicz heart $M^\Phi(\mu)$ for finite measures μ on $\mathcal{A}$ and $\mathbb{R}_+$-valued Young functions Φ (Theorem 5.16). To this end, recall our measurable space $(\Omega, \mathcal{A})$, and let $\mu \neq 0$ be a finite measure on $\mathcal{A}$.

A function $\Phi\colon \mathbb{R} \to \overline{\mathbb{R}}_+$ is called a Young function, if it is convex, even, lower semicontinuous and non-trivial (i.e. $\Phi(x) \in \mathbb{R}_+$ for some $x > 0$) with $\Phi(0) = 0$ and $\lim_{x\to\infty} \Phi(x) = \infty$. Given a Young function Φ, the convex conjugate (also referred to as Fenchel conjugate, or Legendre–Fenchel conjugate) is defined by $\Psi(y) = \sup_{x\geq 0}(x|y| - \Phi(x))$ for every $y \in \mathbb{R}$. Since the operation of convex conjugation preserves both convexity and lower semicontinuity (Schmock, 2024, Remark 14.23), it follows that Ψ is a Young function

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

too, and we call (Φ, Ψ) a complementary pair of Young functions.

Every $\mathbb{R}_+$-valued Young function Ψ is continuous and admits an integral representation, $\Phi(x) = \int_0^{|x|} \varphi(t)\,\mathrm{d}t$, where $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing and left-continuous with $\varphi(0) = 0$ (Rao and Ren, 1991, Corollary 1.3.2). In addition, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, then for any random variable $\xi \geq 0$, an application of Fubini's theorem shows that

$$\mathbb{E}[\Phi(\xi)] = \int_0^\infty \mathbb{P}(\xi > t)\varphi(t)\,\mathrm{d}t = \int_0^\infty \mathbb{P}(\xi \geq t)\varphi(t)\,\mathrm{d}t,$$

which partially generalizes Kallenberg (2021, Lemma 4.4) from $\Phi(x) = |x|^p$ for $p \geq 1$ to all $\mathbb{R}_+$-valued Young functions Φ such as, for example, $\Phi(x) = \mathrm{e}^{x^2} - 1$, where $\varphi(t) = 2t\mathrm{e}^{t^2}$.

Given a complementary pair (Φ, Ψ) of Young functions, $\mathcal{L}^\Phi(\mu)$ consists of all $\mathcal{A}$-measurable $f\colon X \to \mathbb{R}$, such that $\int_X \Phi(\alpha f)\,\mathrm{d}\mu < \infty$ for some $\alpha > 0$, and the Orlicz space $L^\Phi(\mu)$ consists of all equivalence classes of functions from $\mathcal{L}^\Phi(\mu)$, where we identify f and g if $f - g = 0$ up to a set of μ-measure zero. Let $\mathcal{A}^\Psi$ denote the set of all $g \in L^\Psi(\mu)$ such that $\int_X \Psi(g)\,\mathrm{d}\mu \leq 1$. We then define the Orlicz norm $\|\cdot\|_\Phi$ on $L^\Phi(\mu)$ as

$$\|f\|_\Phi = \sup_{g \in \mathcal{A}^\Psi} \int_X |fg|\,\mathrm{d}\mu,$$

and note that $(L^\Phi(\mu), \|\cdot\|_\Phi)$ is a Banach space (Rao and Ren, 1991, Proposition 3.3.11). Let us mention that besides $\|\cdot\|_\Phi$, there exists an equivalent norm on $L^\Phi(\mu)$ that is also popular in the literature, the Luxemburg norm; we refer to Rao and Ren (1991, Section 3.2) for details. The set $\mathcal{M}^\Phi(\mu)$ consists of all $\mathcal{A}$-measurable $f\colon X \to \mathbb{R}$, such that $\int_X \Phi(\alpha f)\,\mathrm{d}\mu < \infty$ for all $\alpha > 0$, and the Orlicz heart $M^\Phi(\mu)$ consists of all equivalence classes of functions from $\mathcal{M}^\Phi(\mu)$, where we again identify f and g if $f - g = 0$ up to a set of μ-measure zero.

A Young function Φ is said to satisfy the Δ_2-condition ($\Phi \in \Delta_2$), if $\Phi(\mathbb{R}) \subset \mathbb{R}_+$, and there exists $x_0 > 0$ and $K > 0$, such that $\Phi(2x) \leq K\Phi(x)$ for all $x \geq x_0$. In that case, $M^\Phi(\mu) = L^\Phi(\mu)$, and simple functions are dense in $L^\Phi(\mu)$ (Rao and Ren, 1991, Corollary 3.4.5). In particular, if $\Phi \in \Delta_2$, then $(M^\Phi(\mu), \|\cdot\|_\Phi)$ is a Banach space. Some (negative) examples of such Young functions Φ are:

- For each $p \in [1, \infty)$ and $a > 0$, the function $\Phi(x) = a|x|^p$ satisfies $\Phi \in \Delta_2$.
- The function $\Phi(x) = \mathrm{e}^{|x|} - 1$ does not satisfy Δ_2, i.e. $\Phi \notin \Delta_2$.
- The function $\Phi(x) = x^2/\log(\mathrm{e} + |x|)$ satisfies $\Phi \in \Delta_2$.
- The function $\Phi(x) = (1 + |x|)\log(1 + |x|) - |x|$ satisfies $\Phi \in \Delta_2$.

Lemma 5.9. *For every Young function Φ, we have $L^\Phi(\mu) \subset L^1(\mu)$.*

Proof. We first claim that there exist $a > 0$ and $b \geq 0$, such that $\Phi(x) \geq ax - b$ for every $x \in \mathbb{R}_+$. To see this, note that the definition of the convex conjugate Ψ implies

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the Fenchel–Young inequality: $\Phi(x) + \Psi(y) \geq xy$ for all $x, y \in \mathbb{R}_+$. There are now two possible cases to consider:

(a) There exists $y^* > 0$ such that $\Psi(y^*) < \infty$. Then $\Phi(x) \geq y^* x - \Psi(y^*)$ for every $x \in \mathbb{R}_+$ (regardless of whether $\Phi(x) \in \mathbb{R}_+$ or $\Phi(x) = \infty$), and setting $a = y^*$ and $b = \Psi(y^*)$ gives the claim.

(b) The effective domain of Ψ is a singleton, i.e. $\{y \in \mathbb{R} \colon \Psi(y) < \infty\} = \{0\}$. Consider the function $f \colon \mathbb{R} \to \mathbb{R}_+$ given by $f(x) = 0$ for every $x \in \mathbb{R}$. Then f is both convex and lower semicontinuous, and the convex conjugate of f is Ψ. The Fenchel–Moreau theorem (Schmock, 2024, Theorem 14.24) states that a function is convex and lower semicontinuous if an only if it is equal to its convex biconjugate, i.e. the convex conjugate of its convex conjugate. But this implies that the only convex and lower semicontinuous function whose convex conjugate is Ψ is given by f. Since f is not a Young function, we can disregard this case.

For $f \in L^\Phi(\mu)$, let $c > 0$ be such that $\int_X \Phi(cf)\,\mathrm{d}\mu < \infty$. Then, by the claimed inequality, $ac|f| \leq \Phi(c|f|) + b = \Phi(cf) + b$, hence

$$\|f\|_1 \leq \frac{1}{ac}\Big(\int_X \Phi(cf)\,\mathrm{d}\mu + b\mu(X)\Big),$$

where the right-hand side is finite. □

We next recall a result that will be needed later, which helps to establish convergence with respect to $\|\cdot\|_\Phi$ (Rao and Ren, 1991, Theorem 3.4.12).

Theorem 5.10. *Let $\Phi \in \Delta_2$, and $(f_n)_{n\in\mathbb{N}}$ be a sequence in $L^\Phi(\mu)$ and $f \in L^\Phi(\mu)$. Then,*

$$\|f_n - f\|_\Phi \to 0 \quad \Longleftrightarrow \quad \int_X \Phi(f_n - f)\,\mathrm{d}\mu \to 0,$$

i.e. there is equivalence between norm convergence and mean convergence. A sufficient condition for mean convergence is if $f_n \to f$ in measure and $\int_X \Phi(f_n)\,\mathrm{d}\mu \to \int_X \Phi(f)\,\mathrm{d}\mu$.

Recall that for $f \in L^2(\mu)$ and a sub-σ-algebra $\tilde{\mathcal{A}} \subset \mathcal{A}$, the conditional expectation $\mathbb{E}_\mu[f \,|\, \tilde{\mathcal{A}}]$ is defined as the orthogonal projection in $L^2(\mu)$ of f onto the closed subspace of all $g \in L^2(\mu)$ that are $\tilde{\mathcal{A}}$-measurable (Schilling, 2017, Definition 27.3). Moreover, for every $p \in [1, \infty]$, the operator $L^2(\mu) \ni f \mapsto \mathbb{E}_\mu[f \,|\, \tilde{\mathcal{A}}]$ has an extension mapping $L^p(\mu)$ onto the subspace of all $g \in L^p(\mu)$ that are $\tilde{\mathcal{A}}$-measurable (Schilling, 2017, Theorem 27.5). In particular, by choosing $p = 1$, we see that for every Young function Φ, every sub-σ-algebra $\tilde{\mathcal{A}} \subset \mathcal{A}$, and every $f \in L^\Phi(\mu)$, a combination of Lemma 5.9 and Schilling (2017, Theorem 27.5) implies that the conditional expectation $\mathbb{E}_\mu[f \,|\, \tilde{\mathcal{A}}]$ is well defined.

With the help of Theorem 5.10, we can formulate and prove a convergence theorem in Orlicz spaces for conditional expectations. This generalizes Kallenberg (2021, Theorem 9.24) from L^1-spaces over probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ to Orlicz spaces L^Φ over

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

finite measure spaces $(X, \mathcal{A}, \mu)$. See also Krickeberg (1964, Theorem 2) for a further generalization to convergence of nets in Orlicz spaces L^Φ over probability spaces.

Theorem 5.11. *Let $(\mathcal{A}_n)_{n\in\mathbb{N}}$ be a filtration of $\mathcal{A}$ and $\mathcal{A}_\infty \coloneqq \bigvee_{n\in\mathbb{N}} \mathcal{A}_n$. If $\Phi \in \Delta_2$, then for every $g \in L^\Phi(\mu)$, we have*

$$\mathbb{E}_\mu[g \,|\, \mathcal{A}_n] \to \mathbb{E}_\mu[g \,|\, \mathcal{A}_\infty], \quad n \to \infty,$$

where convergence holds μ-almost everywhere and in $L^\Phi(\mu)$.

Proof. Let $c > 0$ be such that $\int_X \Phi(cg)\,\mathrm{d}\mu < \infty$, and let $M_n = \mathbb{E}_\mu[g \,|\, \mathcal{A}_n]$ for $n \in \mathbb{N}\cup\{\infty\}$. Then, for every $n \in \mathbb{N} \cup \{\infty\}$, M_n is well defined by Lemma 5.9 and Schilling (2017, Theorem 27.5). By the conditional Jensen inequality (Schilling, 2017, Theorem 27.16) and the elementary properties of the conditional expectation (Schilling, 2017, Theorem 27.11),

$$\int_X \Phi(cM_n)\,\mathrm{d}\mu \le \int_X \mathbb{E}_\mu[\Phi(cg) \,|\, \mathcal{A}_n]\,\mathrm{d}\mu = \int_X \Phi(cg)\,\mathrm{d}\mu < \infty,$$

which shows that $M_n \in L^\Phi(\mu)$ for every $n \in \mathbb{N} \cup \{\infty\}$.

By Doob's martingale convergence theorem (Schilling, 2017, Theorem 27.19(i)), M converges to M_∞ both μ-almost everywhere, and in $L^1(\mu)$. According to Theorem 5.10, convergence in norm follows if we can show that $\int_X \Phi(M_n)\,\mathrm{d}\mu$ converges to $\int_X \Phi(M_\infty)\,\mathrm{d}\mu$. By Vitali's convergence theorem (Schmock, 2024, Theorem 4.60), this is certainly the case if the set $\{\Phi(M_n)\colon n \in \mathbb{N}\}$ is uniformly integrable.

Uniform integrability is implied by the condition (Schilling, 2017, Theorem 22.9(ix))

$$\lim_{c\to\infty} \sup_{n\in\mathbb{N}} \int_{\{\Phi(M_n)>c\}} \Phi(M_n)\,\mathrm{d}\mu = 0.$$

By Markov's inequality and Jensen's inequality for conditional expectations,

$$\mu(\Phi(M_n) > c) \le \frac{1}{c}\int_X \Phi(M_n)\,\mathrm{d}\mu \le \frac{1}{c}\int_X \Phi(g)\,\mathrm{d}\mu.$$

Since $\Phi \in \Delta_2$, we have that $L^\Phi(\mu) = M^\Phi(\mu)$, and therefore $\int_X \Phi(g)\,\mathrm{d}\mu$ is finite. This shows that $\mu(\Phi(M_n) > c)$ converges to 0 as $c \to \infty$, uniformly in n.

Since $\Phi(g) \in L^1(\mu)$, the measure $\nu(A) \coloneqq \int_A \Phi(g)\,\mathrm{d}\mu$ is absolutely continuous with respect to μ on $\mathcal{A}$ with Radon–Nikodým density $\Phi(g)$. By Williams (1991, Lemma 13.1(a)), there exists for each $\varepsilon > 0$ some $\delta > 0$, such that $\mu(A) < \delta$ for $A \in \mathcal{A}$ implies $\nu(A) < \varepsilon$. Let $c^* > 0$ be such that $\mu(\Phi(M_n) > c^*) < \delta$ for all $n \in \mathbb{N}$. Then, for every $c \ge c^*$,

$$\sup_{n\in\mathbb{N}} \int_{\{\Phi(M_n)>c\}} \Phi(M_n)\,\mathrm{d}\mu \le \sup_{n\in\mathbb{N}} \int_{\{\Phi(M_n)>c\}} \Phi(g)\,\mathrm{d}\mu = \sup_{n\in\mathbb{N}} \nu(\Phi(M_n) > c) \le \varepsilon.$$

This shows that $\{\Phi(M_n)\colon n \in \mathbb{N}\}$ is uniformly integrable, which concludes the proof. □

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Before we finally formulate and prove the universal approximation theorem in Orlicz spaces, we recall the following representation of the continuous dual spaces of $M^\Phi(\mu)$ and of $L^\Phi(\mu)$ (Rao and Ren, 1991, Theorem 4.1.6).

Theorem 5.12. *Let (Φ, Ψ) be a complementary pair of Young functions. To each $F \in (M^\Phi(\mu))^*$ there corresponds a unique $g \in L^\Psi(\mu)$, such that*

$$F(f) = \int_X fg \, \mathrm{d}\mu, \quad f \in M^\Phi(\mu),$$

and $\|F\|_{\mathrm{op}} = \|g\|_\Psi$, where $\|F\|_{\mathrm{op}}$ denotes the operator norm of F. Thus, $(M^\Phi(\mu))^*$ is isometrically isomorphic to $L^\Psi(\mu)$, i.e. $(M^\Phi(\mu))^* \simeq L^\Psi(\mu)$. Moreover, if $\Phi \in \Delta_2$, then $(L^\Phi(\mu))^* \simeq L^\Psi(\mu)$.

General Assumption 5.13. For the remainder of this section, let us assume that

(a) X is a locally convex Hausdorff topological vector space, $\Gamma = X^*$ and $\mathcal{A} = \sigma(\Gamma)$.

(b) The activation function $\psi\colon \mathbb{R} \to \mathbb{R}$ is bounded, measurable and sigmoidal, meaning that $\lim_{x\to-\infty} \psi(x) = 0$, and $\lim_{x\to+\infty} \psi(x) = 1$.

Remark 5.14. Assumption 5.13(b) follows Cybenko (1989). The conditions that Hornik (1991) imposed on ψ are less restrictive, requiring ψ to be non-constant in place of sigmoidal. We leave a potential relaxation of Assumption 5.13(b) to future work.

The following proposition complements Cybenko (1989, Lemma 1) in two ways. It extends the setting from signed measures on $\mathcal{B}_{\mathbb{R}^d}$ with $d \in \mathbb{N}$ to signed measures on $\sigma(\Gamma)$ for general locally convex Hausdorff topological vector spaces X with $\Gamma = X^*$. Moreover, we provide an alternative proof using a monotone class-type argument.

Proposition 5.15. *The activation function ψ is discriminatory, meaning that there cannot exist a real-valued signed measure $\nu \neq 0$ on $\mathcal{A}$ such that $\int_X \psi(f(x) + \beta)\, \nu(\mathrm{d}x) = 0$ for every $f \in \Gamma$ and $\beta \in \mathbb{R}$.*

Proof. Let us assume to the contrary that there exists such a signed measure, and let $\nu = \nu^+ - \nu^-$ be the Hahn–Jordan decomposition of ν into two finite and mutually singular measures on $\mathcal{A}$. We aim to show that $\nu^+ = \nu^-$. By (Bogachev, 2007, Lemma 7.13.5), this is certainly the case if the characteristic functions $\widetilde{\nu}^+$, $\widetilde{\nu}^-$ of ν^+, ν^- agree on Γ. Given $f \in \Gamma$, an application of Euler's formula shows that

$$\widetilde{\nu}^+(f) - \widetilde{\nu}^-(f) = \int_X \exp(\mathrm{i} f(x))\, \nu(\mathrm{d}x) = \int_X \cos(f(x))\, \nu(\mathrm{d}x) + \mathrm{i} \int_X \sin(f(x))\, \nu(\mathrm{d}x), \quad (5.4)$$

so we must show that both integrals on the right-hand side of Eq. (5.4) are zero. Note that both the sine and cosine function are bounded, Borel-measurable functions.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

We aim to invoke the monotone class theorem (Schmock, 2024, Theorem 15.69). To this end, let $\mathcal{H}$ denote the set of all bounded, Borel-measurable functions $h\colon \mathbb{R} \to \mathbb{R}$ such that $\int_X h(f(x))\,\nu(\mathrm{d}x) = 0$ for all $f \in \Gamma$.

(a) By construction, $\mathcal{H}$ is a real vector space.
(b) The constant function $\mathbb{1}_{\mathbb{R}}$ is contained in $\mathcal{H}$, since for every $f \in \Gamma$,

$$\int_X \mathbb{1}_{\mathbb{R}}(f(x))\,\nu(\mathrm{d}x) = \nu(X) = \frac{1}{\psi(\beta)} \int_X \psi(0+\beta)\,\nu(\mathrm{d}x) = 0,$$

where β is any real constant such that $\psi(\beta) \neq 0$ (which we can always find since ψ is sigmoidal), and $f = 0$ is trivially contained in Γ.
(c) If $(h_n)_{n\in\mathbb{N}}$ is a pointwise increasing sequence of $\mathbb{R}_+$-valued functions in $\mathcal{H}$ that is uniformly bounded by a constant $C \in \mathbb{R}_+$, then the pointwise limit function is also contained in $\mathcal{H}$ by the dominated convergence theorem applied to ν^+ and ν^-.

Let $\mathcal{E}$ denote the intersection-stable system of half-open intervals of the form $(a, b]$ for $-\infty < a \leq b < \infty$, where we convene that $(a, b] = \varnothing$ for $a = b$. Then $\mathcal{E}$ generates $\mathcal{B}_{\mathbb{R}}$, i.e. $\mathcal{B}_{\mathbb{R}} = \sigma(\mathcal{E})$. If we can show that $\mathbb{1}_E \in \mathcal{H}$ for every $E \in \mathcal{E}$, then we are done. Fix $\theta \in \mathbb{R}$, and let $(s_n)_{n\in\mathbb{N}}$ be an increasing sequence of positive scalars with $\lim_{n\to\infty} s_n = \infty$. Using the assumption on Γ and the fact that ψ is sigmoidal, an application of the dominated convergence theorem yields

$$\begin{aligned}\int_X \mathbb{1}_{(\theta,\infty)}(f(x))\,\nu(\mathrm{d}x) &= \nu(f(x) > \theta) \\ &= \lim_{\gamma\to-\infty} \left(\nu(f(x) > \theta) + \psi(\gamma)\nu(f(x) = \theta)\right) \\ &= \lim_{\gamma\to-\infty} \lim_{n\to\infty} \int_X \psi(s_n(f(x) - \theta) + \gamma)\,\nu(\mathrm{d}x) = 0,\end{aligned}$$

for all $f \in \Gamma$. By taking linear combinations of $\mathbb{1}_{(\theta,\infty)}$ for $\theta \in \mathbb{R}$, we see that $\mathbb{1}_E \in \mathcal{H}$ for every $E \in \mathcal{E}$. The monotone class theorem (Schmock, 2024, Theorem 15.69) now shows that $\mathcal{H}$ contains all bounded, $\sigma(\mathcal{E}) = \mathcal{B}_{\mathbb{R}}$-measurable functions $h\colon \mathbb{R} \to \mathbb{R}$, and in particular the sine and cosine function. But in this case, the right-hand side of Eq. (5.4) is zero for each $f \in \Gamma$, which shows that ν^+ must be equal to ν^-, hence $\nu \equiv 0$. □

The following theorem generalizes Theorem 1 in Hornik (1991) in two ways. First, we move from L^p-spaces to the more general class of Orlicz spaces. Moreover, we only require X to be a locally convex Hausdorff topological vector space, rather than the Euclidean space $\mathbb{R}^d$ with $d \in \mathbb{N}$. Recall Standing Assumption 5.13, that $\mu \neq 0$ is assumed to be a finite measure on $\mathcal{A} = \sigma(\Gamma)$, and (Φ, Ψ) be a complementary pair of Young functions.

Theorem 5.16 (Universal approximation in Orlicz spaces)**.** *Let Φ be an $\mathbb{R}_+$-valued Young function. Then the set $\mathcal{NN}(\Gamma, \psi)$ of neural networks is dense in the Orlicz heart*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

$M^\Phi(\mu)$. *Moreover, if in addition* $\Phi \in \Delta_2$, *then* $\mathcal{NN}(\Gamma, \psi)$ *is dense in the Orlicz space* $L^\Phi(\mu)$.

Proof. We claim that $L^\infty(\mu) \subset M^\Phi(\mu)$. To see this, note that Φ is $\mathbb{R}_+$-valued and non-decreasing in the sense that $|x| \leq |y|$ implies $\Phi(x) \leq \Phi(y)$. Therefore, for every $\alpha > 0$, we have that $\Phi(\alpha f)$ is μ-almost everywhere bounded by $\Phi(\alpha \|f\|_\infty)$, hence $\int_X \Phi(\alpha f)\,\mathrm{d}\mu \leq \Phi(\alpha\|f\|_\infty)\mu(X) < \infty$.

Since ψ is bounded, $\mathcal{NN}(\Gamma, \psi) \subset L^\infty(\mu)$ and therefore $\mathcal{NN}(\Gamma, \psi) \subset M^\Phi(\mu)$. If $\mathcal{NN}(\Gamma, \psi)$ were not dense in $M^\Phi(\mu)$ then, as a consequence of the Hahn–Banach theorem (Rudin, 1987, Theorem 5.19), there exists $F \in (M^\Phi(\mu))^*$ such that $F \not\equiv 0$ and $F(f) = 0$ for each $f \in \mathcal{NN}(\Gamma, \psi)$. By Theorem 5.12, there exists a unique $g \in L^\Psi(\mu)$ such that $F(f) = \int_X fg\,\mathrm{d}\mu$ for every $f \in M^\Phi(\mu)$.

Let ν denote the absolutely continuous signed measure $\nu(A) := \int_A g\,\mathrm{d}\mu$ on $\mathcal{A}$. By Hölder's inequality (Rao and Ren, 1991, Proposition 3.3.1), we have for each $A \in \mathcal{A}$,

$$|\nu(A)| \leq \int_X \mathbb{1}_A |g|\,\mathrm{d}\mu \leq \|\mathbb{1}_A\|_\Phi \|g\|_\Psi < \infty,$$

hence ν is a finite signed measure on $\mathcal{A}$. Since $F \not\equiv 0$ and $F(f) = \int_X f(x)\,\nu(\mathrm{d}x)$ it follows that $\nu \not\equiv 0$. Recall that F vanishes on $\mathcal{NN}(\Gamma, \psi)$, hence

$$\int_X \psi(f(x) + \beta)\,\nu(\mathrm{d}x) = 0, \quad f \in \Gamma,\ \beta \in \mathbb{R}. \tag{5.5}$$

Proposition 5.15 shows that ψ is discriminatory, hence (5.5) implies that $\nu \equiv 0$, which further implies that $F \equiv 0$, yielding a contradiction.

Finally, if $\Phi \in \Delta_2$, then $L^\Phi(\mu) = M^\Phi(\mu)$, and $\mathcal{NN}(\Gamma, \psi)$ is dense in $L^\Phi(\mu)$ as well. □

In this section, we have seen that the approximation capabilities of neural networks extend to very general spaces. In particular, we have studied neural networks on general locally convex Hausdorff topological vector spaces and introduced several abstract examples. Note that we have only covered the case of approximating real-valued functions; we leave the extension to higher- or infinite-dimensional codomains to future work. Likewise, the motivation for studying approximation capabilities in Orlicz spaces stems from the goal to study optimal algorithmic hedging with respect to risk measures on Orlicz hearts (Cheridito and Li, 2009); while Section 5.3 only covers the case of algorithmic mean-variance hedging, we leave the extension to Orlicz spaces to future work. In Section 5.2 we will see how the results from this section translate to the approximation of optimal trading decisions in financial markets.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

5.2. Algorithmically generated random variables

The results from Section 5.1 have demonstrated useful approximation capabilities of neural networks in very general classes of function spaces. Taking a step towards mathematical finance, we now transfer the results from the previous section to a probabilistic setting that is suitable for studying algorithmic trading strategies in financial markets. This section aims to demonstrate approximation capabilities in spaces of random variables. The key element here will be to impose measurability with respect to a σ-algebra that is generated by a collection of random variables which represent observable market information that can be processed in order to perform informed trading decisions.

Our next result demonstrates that even in case that there is an uncountable amount of random variables that generates the observable information, it is in many cases possible to restrict to the countable case. For an application of the following proposition, it may be useful to recall that in separable metric spaces, the Borel σ-algebra is always countably generated.

Proposition 5.17. *Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be measurable spaces. Let I be a non-empty set, and $(\mathcal{F}_i)_{i\in I}$ be an indexed family of sub-σ-algebras of $\mathcal{F}$. For each $J \subset I$, denote $\mathcal{F}_J = \sigma(\bigcup_{j\in J} \mathcal{F}_j)$. Let $g\colon \Omega \to S$ be an $\mathcal{F}_I$-$\mathcal{S}$-measurable function, such that the trace-σ-algebra of $\mathcal{S}$ on $g(\Omega)$, which we denote $\mathcal{S}_g$, is countably generated. Then there exists a countable set $J \subset I$ such that g is $\mathcal{F}_J$-$\mathcal{S}$-measurable.*

Proof. Without loss of generality, we may assume I to be an uncountable set. Let $(S_n)_{n\in\mathbb{N}}$ be a sequence of elements of $\mathcal{S}$, such that $\mathcal{S}_g = \sigma(S_n \cap g(\Omega)\colon n \in \mathbb{N})$. For every $n \in \mathbb{N}$, let $F_n = g^{-1}(S_n) \in \mathcal{F}_I$, and $\mathcal{F}_g := \sigma(F_n\colon n \in \mathbb{N})$. Then g is $\mathcal{F}_g$-$\mathcal{S}$-measurable, $\mathcal{F}_g$ is countably generated, and $\mathcal{F}_g \subset \mathcal{F}_I$.

Next, we claim that

$$\mathcal{F}_I = \bigcup_{\substack{J\subset I \\ J \text{ countable}}} \mathcal{F}_J. \tag{5.6}$$

To this end, let us first verify that the right-hand side of Eq. (5.6), which we denote $\mathcal{G}$, is indeed a σ-algebra. For every countable $J \subset I$, clearly $\Omega \in \mathcal{F}_J$, and thus $\Omega \in \mathcal{G}$. If $B \in \mathcal{G}$, then there exists a countable $J \subset I$, such that $B \in \mathcal{F}_J$ and thus $B^{\mathsf{c}} \in \mathcal{F}_J$, which implies that $B^{\mathsf{c}} \in \mathcal{G}$. Finally, let $(B_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{G}$. For each $n \in \mathbb{N}$, there exists a countable $J_n \subset I$, such that $B_n \in \mathcal{F}_{J_n}$. The set $J = \bigcup_{n\in\mathbb{N}} J_n$ is countable, and $B_n \in \mathcal{F}_J$ for each $n \in \mathbb{N}$, hence $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{F}_J$ and thus $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{G}$.

This shows that $\mathcal{G}$ is indeed a σ-algebra. In order to show equality in Eq. (5.6), note that $\bigcup_{i\in I} \mathcal{F}_i$ generates $\mathcal{F}_I$. Let $F \in \bigcup_{i\in I} \mathcal{F}_i$. Then, there exists $i \in I$ such that $F \in \mathcal{F}_{\{i\}} \subset \mathcal{G}$, which implies $\mathcal{F}_I \subset \mathcal{G}$. On the other hand, for each countable $J \subset I$, obviously $\mathcal{F}_J \subset \mathcal{F}_I$ and thus $\mathcal{G} \subset \mathcal{F}_I$.

In order to conclude the proof, recall that g is measurable with respect to the countably generated σ-algebra $\mathcal{F}_g = \sigma(F_n\colon n \in \mathbb{N})$, and that $\mathcal{F}_g \subset \mathcal{F}_I$. For each $n \in \mathbb{N}$, Eq. (5.6)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

implies the existence of a countable $J_n \subset I$, such that $F_n \in \mathcal{F}_{J_n}$. Setting $J \coloneqq \bigcup_{n\in\mathbb{N}} J_n$ yields the assertion. □

The proof of Proposition 5.17 used the fact that we can identify g as $\mathcal{F}_g$-measurable, where $\mathcal{F}_g$ is countably generated. If S is a metric space, an alternative argument can be made via simple functions. Recall that a function $f\colon \Omega \to S$ is simple, if $f(\Omega) \subset S$ is a finite set.

Lemma 5.18. *Let $(\Omega, \mathcal{F})$ be a measurable space, (S, ϱ) be a separable metric space and $\mathcal{B}_S$ its Borel σ-algebra. Let $g\colon \Omega \to S$ be measurable. Then, there exists a sequence $(g_n)_{n\in\mathbb{N}}$ of $\mathcal{F}$-$\mathcal{B}_S$-measurable simple functions, such that $\lim_{n\to\infty} g_n(\omega) = g(\omega)$ for every $\omega \in \Omega$. Moreover, there exists a countably generated sub-σ-algebra $\mathcal{F}_g \subset \mathcal{F}$, such that g is $\mathcal{F}_g$-$\mathcal{B}_S$-measurable.*

Proof. Let $\{s_n\colon n \in \mathbb{N}\}$ be a countable dense subset of S, and $S_n = \{s_1, \dots, s_n\}$, $n \in \mathbb{N}$. For every $n \in \mathbb{N}$, let

$$g_n(\omega) = \operatorname*{arg\,min}_{s\in S_n} \varrho(s, g(\omega)), \quad \omega \in \Omega.$$

If several $s \in S_n$ realize the minimum, then we choose the one with the lowest index.

By construction, g_n is simple and $\mathcal{F}$-$\mathcal{B}_S$-measurable. Moreover, for each $\omega \in \Omega$ and $\varepsilon > 0$, there exists $m \in \mathbb{N}$ such that $\varrho(s_m, g(\omega)) < \varepsilon$. But then $\varrho(g_n(\omega), g(\omega)) < \varepsilon$ for every $n \geq m$, which proves the pointwise convergence. Each g_n, being simple, is measurable with respect to a finitely generated σ-algebra $\sigma(F_1^n, \dots, F_{m_n}^n)$ for some sets $F_j^n \in \mathcal{F}$. Since the countable union of finite sets is again countable, we can consider all F_j^n to be contained in one common sequence, which we denote $(F_n)_{n\in\mathbb{N}}$. Since measurability carries over to the pointwise limit (Kallenberg, 2021, Lemma 1.11), g is measurable with respect to the countably generated σ-algebra $\mathcal{F}_g = \sigma(F_n\colon n \in \mathbb{N})$. □

For the remainder of this chapter, let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a complete probability space. Moreover, let $Y = (Y_t)_{t\geq 0}$ be an $\mathbb{R}^d$-valued stochastic process for some dimension $d \in \mathbb{N}$. In analogy to Buehler et al. (2019), the process Y represents all market information. We denote by $\mathbb{F}^0 = (\mathcal{F}_t^0)_{t\geq 0}$ the natural filtration generated by Y, i.e. $\mathcal{F}_t^0 = \sigma(Y_s\colon 0 \leq s \leq t)$ for $t \in \mathbb{R}_+$, and set $\mathcal{F}_\infty^0 = \sigma(\bigcup_{t\geq 0} \mathcal{F}_t^0) = \sigma(Y_s\colon s \in \mathbb{R}_+)$. As it is common in the literature to assume the usual hypotheses to hold, it will also be necessary to consider the enlargement of $\mathbb{F}^0$ with $\mathbb{P}$-null sets. To this end, let $\mathcal{N}_\mathbb{P}$ denote the set of all $\mathbb{P}$-null sets of $\mathcal{F}$. Let $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$, where $\mathcal{F}_t \coloneqq \sigma(\mathcal{N}_\mathbb{P} \cup \mathcal{F}_t^0)$ for every $t \in \mathbb{R}_+$, and $\mathcal{F}_\infty = \sigma(\mathcal{N}_\mathbb{P} \cup \mathcal{F}_\infty^0)$. Then, $\mathbb{F}$ is an enlargement of $\mathbb{F}^0$, and $\mathcal{F}_0$ contains all $\mathbb{P}$-null sets of $\mathcal{F}$.

Remark 5.19. Compared to Proposition 5.17, a somewhat more direct statement can be obtained under additional assumptions (Cohen and Elliott, 2015, Exercise 3.4.13). To this end, assume that Y is $\mathbb{P}$-almost surely càdlàg. For $t > 0$, let $\mathbb{D}_t$ denote a countable dense subset of $[0, t)$ such as, for example, the set of all dyadic rational numbers in $[0, t)$.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In that case,

$$\mathcal{F}_t^0 = \sigma(Y_t) \vee \sigma\Big(\bigcup_{s \in \mathbb{D}_t} \sigma(Y_s) \Big) \vee \mathcal{N}_{\mathbb{P}}. \tag{5.7}$$

Besides representations as in Eq. (5.7), it will be useful to describe σ-algebras at $\mathbb{F}^0$- and at $\mathbb{F}$-stopping times. In the latter case, we will assume Y to be a Lévy process (although stationarity of increments will not be required) in the context of He et al. (1992, Chapter 13.4), while in the former case, we need to impose an additional assumption on Ω, which has been suggested by Shiryaev (2008, Section 1.2), but also appears in He et al. (1992, Problem 3.16) as well as Stroock and Varadhan (2006, Lemma 1.3.3) and Karatzas and Shreve (1998, Lemma 5.4.18) for some special cases.

Definition 5.20. We say that Ω is sufficiently rich, if for each $t \geq 0$ and $\omega \in \Omega$, there exists $\tilde{\omega} \in \Omega$, such that $Y_s(\tilde{\omega}) = Y_{s \wedge t}(\omega)$ for each $s \geq 0$.

Proposition 5.21. *In the context of Proposition 5.17, and for our complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with filtrations $\mathbb{F}^0$ and $\mathbb{F}$,*

(a) Fix $t \in [0, \infty]$, and let $g\colon \Omega \to \mathbb{R}$ be $\mathcal{F}_t^0$-measurable. Then, there exists a countable set $J \subset [0,t] \cap \mathbb{R}_+$ such that, denoting $\mathcal{F}_J = \sigma(Y_t : t \in J)$, g is $\mathcal{F}_J$-measurable.

(b) Assume that Ω is sufficiently rich. Let $\tau\colon \Omega \to \mathbb{R}_+$ be an $\mathbb{F}^0$-stopping time, and let $g\colon \Omega \to \mathbb{R}$ be $\mathcal{F}_\tau$-measurable. Then, there exists a countable set $J \subset \mathbb{R}_+$ such that, denoting $\mathcal{F}_J = \sigma(Y_t^\tau : t \in J)$, g is $\mathcal{F}_J$-measurable.

(c) Assume that Y is a Lévy process. Let $\tau\colon \Omega \to \mathbb{R}_+$ be an $\mathbb{F}$-stopping time, and let $g\colon \Omega \to \mathbb{R}$ be $\mathcal{F}_\tau$-measurable. Then, there exists a countable set $J \subset \mathbb{R}_+$ such that, denoting $\mathcal{F}_J = \sigma(\tau) \vee \sigma(Y_t^\tau : t \in J) \vee \mathcal{N}_{\mathbb{P}}$, g is $\mathcal{F}_J$-measurable.

Proof. The first claim follows from Proposition 5.17 with $I = [0,t] \cap \mathbb{R}_+$ and $\mathcal{F}_i = \sigma(Y_i)$, $i \in I$.

The assumption of Ω being sufficiently rich implies that $\mathcal{F}_\tau = \sigma(Y_t^\tau : t \in \mathbb{R}_+)$ (Shiryaev, 2008, Section 1.2, Theorem 6). The second claim then follows from Proposition 5.17 with $I = \mathbb{R}_+$ and $\mathcal{F}_i = \sigma(Y_i^\tau)$, $i \in I$.

Finally, the assumption of Y being a Lévy process implies that $\mathcal{F}_\tau = \sigma(\tau) \vee \sigma(Y_t^\tau : t \in \mathbb{R}_+) \vee \mathcal{N}_{\mathbb{P}}$ (He et al., 1992, Theorem 13.45). The last claim then follows from Proposition 5.17 with $I = \mathbb{R}_+ \cup \{a, b\}$ and $\mathcal{F}_i = \sigma(Y_i^\tau)$, $i \in \mathbb{R}_+$ as well as $\mathcal{F}_a = \sigma(\tau)$ and $\mathcal{F}_b = \sigma(\mathcal{N}_{\mathbb{P}})$, which concludes the proof. □

Definition 5.22. Motivated by Proposition 5.21, we distinguish the following three cases:

- Let $t \in [0, \infty]$. We call an $\mathcal{F}_t$-measurable $\varphi\colon \Omega \to \mathbb{R}$ algorithmically generated, if there exists $n \in \mathbb{N}$, deterministic times $0 \leq t_1 < \ldots < t_n$ in $[0,t] \cap \mathbb{R}_+$ and a neural

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

network $f \in \mathcal{NN}((\mathbb{R}^{d\times n})^*, \psi)$, such that

$$\varphi = f(Y_{t_1}, \dots, Y_{t_n}). \tag{5.8}$$

- If Ω is sufficiently rich, and $\tau\colon \Omega \to \mathbb{R}_+$ an $\mathbb{F}^0$-stopping time, then we call an $\mathcal{F}_\tau$-measurable $\varphi\colon \Omega \to \mathbb{R}$ algorithmically generated, if there exists $n \in \mathbb{N}$, deterministic times $0 \le t_1 < \ldots < t_n < \infty$ and a neural network $f \in \mathcal{NN}((\mathbb{R}^{d\times n})^*, \psi)$, such that

$$\varphi = f(Y^\tau_{t_1}, \dots, Y^\tau_{t_n}). \tag{5.9}$$

- If Y is a Lévy process, and $\tau\colon \Omega \to \mathbb{R}_+$ an $\mathbb{F}$-stopping time, then we call an $\mathcal{F}_\tau$-measurable $\varphi\colon \Omega \to \mathbb{R}$ algorithmically generated, if there exists $n \in \mathbb{N}$, deterministic times $0 \le t_1 < \ldots < t_n < \infty$ and a neural network $f \in \mathcal{NN}((\mathbb{R}^{1+d\times n})^*, \psi)$, such that

$$\varphi = f(\tau, Y^\tau_{t_1}, \dots, Y^\tau_{t_n}). \tag{5.10}$$

We denote by $\mathcal{NN}_t(\psi)$ and $\mathcal{NN}_\tau(\psi)$ the corresponding sets of algorithmically generated random variables. For notational simplicity, we implicitly distinguish between the latter two cases depending on whether τ is an $\mathbb{F}^0$- or $\mathbb{F}$-stopping time.

Recall that Section 5.1 dealt with approximation capabilities of neural networks in Orlicz spaces. The following theorem combines the universal approximation property established in Theorem 5.16 with the refined treatment of measurability from Proposition 5.21. It extends Øksendal (2003, Lemma 4.3.1). Recall Standing Assumption 5.13(b). For purpose of the next Section 5.3, the last case will be of particular importance.

Theorem 5.23. *Let (Φ, Ψ) be a complementary pair of Young functions, where $\Phi \in \Delta_2$.*

(a) Let I be a non-empty set, and $(Z_i)_{i\in I}$ an indexed family of $\mathcal{F}$-measurable random variables $Z_i\colon \Omega \to \mathbb{R}^d$. Let $\widetilde{\mathcal{F}} = \sigma(\mathcal{N}_\mathbb{P} \cup \sigma(Z_i\colon i \in I))$, and $g \in \mathcal{L}^\Phi(\mathbb{P})$ be $\widetilde{\mathcal{F}}$-measurable. Then, for every $\varepsilon > 0$, there exists $n \in \mathbb{N}$, a finite set $J = \{j_1, \dots, j_n\} \subset I$, and a neural network $f \in \mathcal{NN}((\mathbb{R}^{d\times n})^*, \psi)$, such that

$$\|g - f(Z_{j_1}, \dots, Z_{j_n})\|_\Phi < \varepsilon.$$

(b) Let $t \in [0, \infty]$ and $g \in \mathcal{L}^\Phi(\mathbb{P})$ be $\mathcal{F}_t$-measurable. Then, for every $\varepsilon > 0$, there exists $f \in \mathcal{NN}_t(\psi)$, such that $\|g - f\|_\Phi < \varepsilon$.

(c) Assume that Ω is sufficiently rich. Let $\tau\colon \Omega \to \mathbb{R}_+$ be an $\mathbb{F}^0$-stopping time, and $g \in \mathcal{L}^\Phi(\mathbb{P})$ be $\mathcal{F}_\tau$-measurable. Then, for every $\varepsilon > 0$, there exists $f \in \mathcal{NN}_\tau(\psi)$, such that $\|g - f\|_\Phi < \varepsilon$.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

(d) Assume that Y is a Lévy process. Let $\tau\colon \Omega \to \mathbb{R}_+$ be an $\mathbb{F}$-stopping time, and $g \in \mathcal{L}^\Phi(\mathbb{P})$ be $\mathcal{F}_\tau$-measurable. Then, for every $\varepsilon > 0$, there exists $f \in \mathcal{NN}_\tau(\psi)$, such that $\|g - f\|_\Phi < \varepsilon$.

Proof. In order to show the first part, note that $g = \mathbb{E}[g \,|\, \widetilde{\mathcal{F}}] = \mathbb{E}[g \,|\, \sigma(Z_i\colon i \in I)] =: \widetilde{g}$, where the equalities are to be understood $\mathbb{P}$-almost surely. Since $\widetilde{g}$ is $\sigma(Z_i\colon i \in I)$-measurable, an application of Proposition 5.17 shows that there exists a countable set $J = \{j_n\colon n \in \mathbb{N}\} \subset I$ such that, denoting $\mathcal{F}_J = \sigma(Z_{j_n}\colon n \in \mathbb{N})$, we have that $\widetilde{g}$ is $\mathcal{F}_J$-measurable, whence $\widetilde{g} = \mathbb{E}[\widetilde{g} \,|\, \mathcal{F}_J]$.

For each $n \in \mathbb{N}$, let $\mathcal{F}_n = \sigma(Z_{j_1}, \dots, Z_{j_n})$. Then, $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is a filtration of $\mathcal{F}_J$ with $\bigcup_{n\in\mathbb{N}} \mathcal{F}_n = \mathcal{F}_J$. Consider the process $M = (M_n)_{n\in\mathbb{N}}$ given by $M_n = \mathbb{E}[\widetilde{g} \,|\, \mathcal{F}_n]$ for $n \in \mathbb{N}$. By Theorem 5.11, M converges to $\mathbb{E}[\widetilde{g} \,|\, \mathcal{F}_J] = \widetilde{g}$ both $\mathbb{P}$-almost surely and in $L^\Phi(\mathbb{P})$. Let $n \in \mathbb{N}$ be large enough, such that $\|g - M_n\|_\Phi = \|\widetilde{g} - M_n\|_\Phi < \varepsilon/2$. By construction, M_n is $\mathcal{F}_n$-measurable. Therefore, by the Doob–Dynkin factorization lemma (Kallenberg, 2021, Lemma 1.14), $M_n = f(Z_{j_1}, \dots, Z_{j_n})$ for some Borel-measurable function $f\colon \mathbb{R}^{dn} \to \mathbb{R}$, where we identify $\mathbb{R}^{d\times n} \simeq \mathbb{R}^{dn}$.

Let μ denote the Borel probability measure that is induced by $(Z_{j_1}, \dots, Z_{j_n})$ on $\mathbb{R}^{dn}$. Since $M_n \in L^\Phi(\mathbb{P})$, it follows that $f \in L^\Phi(\mu)$. By Theorem 5.16, there exists a sequence $(f_m)_{m\in\mathbb{N}}$ of neural networks that converges to f in $L^\Phi(\mu)$. Now choose $m \in \mathbb{N}$ large enough such that $\|f - f_m\|_\Phi < \varepsilon/2$. After applying the triangle inequality, we obtain

$$\|g - f_m(Z_{j_1}, \dots, Z_{j_n})\|_\Phi \le \|g - M_n\|_\Phi + \|f - f_m\|_\Phi < \varepsilon,$$

which concludes the proof for the first case.

In using Proposition 5.21(b), the second part can be shown precisely as the first part, with the indexed family $(Y_s)_{s\in I}$ and $I = [0,t] \cap \mathbb{R}_+$. Using Proposition 5.21(c), the third part can also be shown precisely as the first part, with the indexed family $(Y_s^\tau)_{s\in I}$ and $I = \mathbb{R}_+$. Finally, the last part can be shown using Proposition 5.21(c) using the family $\{Y_s\colon s \in \mathbb{R}_+\} \cup \{\tau\}$. □

Remark 5.24. Recall that we assumed $(\Omega, \mathcal{F}, \mathbb{P})$ to be a complete probability space, and $\mathcal{F}_0$ contains all $\mathbb{P}$-null sets of $\mathcal{F}$. We aim to work with the filtered complete probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ under the usual hypotheses. In order for this to work, we need to ensure $\mathbb{F}$ to be right-continuous. We are therefore looking for sufficient conditions which turn $\mathbb{F}$ into a right-continuous filtration of $\mathcal{F}$. According to Schmock (2024, Remark 3.23), the filtration $\mathbb{F}$ is right-continuous, provided that Y is right-continuous w.r.t. the Euclidean norm on $\mathbb{R}^d$, and Y has independent future increments w.r.t. $\mathbb{F}^0$. This means that Brownian motion and, more generally, Lévy processes are admissible processes for Y.

Motivated by Remark 5.24 and Proposition 5.21(d), we introduce the

General Assumption 5.25. The process Y is a Lévy process, which implies that the filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ is complete and satisfies the usual hypotheses.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In what follows, we denote by $\mathcal{M}(\psi)$ the set of all real-valued $(\mathbb{F},\mathbb{P})$-martingales M that are closable by some $\varphi \in \mathcal{NN}_\infty(\psi)$. Due to Standing Assumption 5.25, every $(\mathbb{F},\mathbb{P})$-martingale has a càdlàg modification (Cohen and Elliott, 2015, Corollary 5.1.9), and we shall always implicitly consider such a version when speaking of an $(\mathbb{F},\mathbb{P})$-martingale.

Corollary 5.26. *In the setting of Example 5.8 and Theorem 5.23, for every $p \in (1,\infty)$, the set $\mathcal{M}(\psi)$ is dense in $\mathcal{H}^p$.*

Proof. First, we will show that $\mathcal{M}(\psi) \subset \mathcal{H}^p$. Let $M \in \mathcal{M}(\psi)$. Then M is closable by some $\varphi \in \mathcal{NN}_\infty(\psi)$. By definition, φ is bounded (because the activation function ψ is bounded), hence $\varphi \in L^p(\mathbb{P})$. But since $\mathcal{H}^p$ can be identified with those elements from $L^p(\mathbb{P})$ that are $\mathcal{F}_\infty$-measurable, we have $M \in \mathcal{H}^p$.

Next, let $N \in \mathcal{H}^p$, and let N_∞ denote its $\mathbb{P}$-almost surely existing limit in $L^p(\mathbb{P})$. Then, N_∞ is $\mathcal{F}_\infty$-measurable, and by Theorem 5.23(b), for every $\varepsilon > 0$, there exists $\varphi \in \mathcal{NN}_\infty(\psi)$, such that $\|N_\infty - \varphi\|_p < \varepsilon/q$, where $1/p + 1/q = 1$. Let $M \in \mathcal{M}(\psi)$ be closable by φ, such that $M_\infty = \mathbb{E}[\varphi \,|\, \mathcal{F}_\infty] = \varphi$. An application of Cohen and Elliott (2015, Lemma 10.1.3) shows that

$$\|N - M\|_{\mathcal{H}^p} \leq q\|N_\infty - M_\infty\|_p < \varepsilon,$$

which concludes the proof. □

5.3. Applications in mathematical finance

5.3.1. Universal approximation of stochastic integral processes

Definition 5.27. Let us define some processes.

- A stochastic process V is said to be simple, if V has a representation

$$V_t = \varphi_0 \mathbb{1}_{\{0\}}(t) + \sum_{i=1}^{n} \varphi_i \mathbb{1}_{(\tau_i,\tau_{i+1}]}(t), \tag{5.11}$$

  where $0 = \tau_1 \leq \ldots \leq \tau_{n+1} < \infty$ is a finite sequence of $\mathbb{F}$-stopping times, and $\varphi_i \in \mathcal{L}^\infty(\mathcal{F}_{\tau_i},\mathbb{P})$ for $i = 0,1,\ldots,n$. We denote by $\mathcal{S}$ the set of all simple processes, and call elements of $\mathcal{S}$ simple strategies.

- A stochastic process V is said to be elementary, if V has a representation (5.11), where the stopping times τ_i are deterministic. We denote by $\mathcal{E}$ the set of all elementary processes, and call elements of $\mathcal{E}$ elementary strategies.

- We denote by $\mathcal{S}(\psi)$ the set of all simple strategies $V \in \mathcal{S}$ that admit a representation (5.11), where $\varphi_i \in \mathcal{NN}_{\tau_i}(\psi)$ for $i = 0,1,\ldots,n$, and call elements of $\mathcal{S}(\psi)$ simple algorithmic strategies.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- We denote by $\mathcal{E}(\psi)$ the set of all elementary strategies $V \in \mathcal{E}$ that admit a representation (5.11), where $\varphi_i \in \mathcal{NN}_{\tau_i}(\psi)$ for $i = 0, 1, \dots, n$, and call elements of $\mathcal{E}(\psi)$ elementary algorithmic strategies.

Proposition 5.28. *Let (Φ, Ψ) be a complementary pair of Young functions, where $\Phi \in \Delta_2$. Then, for each $V \in \mathcal{E}$ and $\varepsilon > 0$, there exists $U \in \mathcal{E}(\psi)$, such that $\|(V-U)^*_\infty\|_\Phi < \varepsilon$. In particular, $\mathcal{E}(\psi)$ is dense in $\mathcal{E}$ for the topology of ucp-convergence. Moreover, for each $V \in \mathcal{S}$ and $\varepsilon > 0$, there exists $U \in \mathcal{S}(\psi)$, such that $\|(V-U)^*_\infty\|_\Phi < \varepsilon$. In particular, $\mathcal{S}(\psi)$ is dense in $\mathcal{S}$ for the topology of ucp-convergence.*

Proof. We argue both cases simultaneously. Let V be given by a representation (5.11), either with deterministic times τ_i or with $\mathbb{F}$-stopping times τ_i. Fix $i \in \{0, 1, \dots, n\}$, and note that $\varphi_i \in \mathcal{L}^\infty(\mathbb{P}) \subset \mathcal{L}^\Phi(\mathbb{P})$. By Theorem 5.23(b) and (d), there exists a sequence $(\varphi_{i,m})_{m\in\mathbb{N}}$ of elements in $\mathcal{NN}_{\tau_i}(\psi)$ that converges to φ_i with respect to $\|\cdot\|_\Phi$.

Let $(V^m)_{m\in\mathbb{N}}$ be the sequence given by

$$V^m = \varphi_{0,m}\mathbb{1}_{\{0\}}(\cdot) + \sum_{i=1}^n \varphi_{i,m}\mathbb{1}_{(\tau_i,\tau_{i+1}]}(\cdot), \quad m \in \mathbb{N}.$$

By construction, it holds that $(V - V^m)^*_\infty \le \sum_{i=0}^n |\varphi_i - \varphi_{i,m}|$. By Rao and Ren (1991, Proposition 3.3.4), the Orlicz norm $\|\cdot\|_\Phi$ is monotone in the sense that $0 \le f_1 \le f_2$ implies $\|f_1\|_\Phi \le \|f_2\|_\Phi$, hence

$$\|(V - V^m)^*_\infty\|_\Phi \le \Big\| \sum_{i=0}^n |\varphi_i - \varphi_{i,m}| \Big\|_\Phi \le \sum_{i=0}^n \|\varphi_i - \varphi_{i,m}\|_\Phi,$$

where the right-hand side converges to zero as $m \to \infty$. Therefore, we can find $m \in \mathbb{N}$ such that $\|(V - V^m)^*_\infty\|_\Phi < \varepsilon$, and the required process U is found by setting $U = V^m$.

In order to prove the last claim, assume Φ to be strictly increasing, e.g. $\Phi(x) = x^p$ for $p \in [1, \infty)$, or $\Phi(x) = x^2/\log(\mathrm{e} + x)$. An application of Chebychev's inequality (Schmock, 2024, Lemma 16.1) yields for every $\varepsilon > 0$,

$$\mathbb{P}\big((V - V^m)^*_\infty > \varepsilon\big) \le \frac{1}{\Phi(\varepsilon)} \int_\Omega \Phi\big((V - V^m)^*_\infty\big)\, \mathrm{d}\mathbb{P}. \tag{5.12}$$

We have already shown that $\|(V - V^m)^*_\infty\|_\Phi$ converges to zero as $m \to \infty$. In recalling that norm convergence implies mean convergence by Theorem 5.10, if follows that the right-hand side of (5.12) converges to zero as $m \to \infty$. Therefore, V^m converges to V uniformly on $\mathbb{R}_+$ in probability, which concludes the proof. □

Elements of $\mathcal{S}(\psi)$ – or, more generally, in $\mathcal{S}$ – are simple enough to allow for a canonical definition of a stochastic integral. Denote by $(\mathcal{L}, \mathcal{T}_{\mathrm{ucp}})$ resp. $(\mathcal{D}, \mathcal{T}_{\mathrm{ucp}})$ the spaces of real-valued càglàd resp. càdlàg $\mathbb{F}$-adapted processes endowed with the topology of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

ucp-convergence. For any $X \in \mathcal{D}$ and $V \in \mathcal{S}(\psi)$, we set

$$I(V,X) = \int V \,\mathrm{d}X \coloneqq f_0(\tau_0, Y_0^{\tau_0})X_0 + \sum_{i=1}^{n} f_i(\tau_i, Y_{t_0^i}^{\tau_i}, \ldots, Y_{t_{m_i}^i}^{\tau_i})(X_{\tau_{i+1}} - X_{\tau_i}),$$

where $f_0, \ldots, f_n$ are neural networks. The corresponding process in $\mathcal{D}$ is then given by the mapping $\mathbb{R}_+ \ni t \mapsto I(V, X^t)$. If X is a semimartingale, then a stochastic integral $\int V \,\mathrm{d}X$ can be defined for a more general class of processes V, namely for all $V \in \mathcal{L}$, see e.g. Protter (2005, Chapter II.4).

It is now a consequence of the Bichteler–Dellacherie theorem (Bichteler, 1979; Dellacherie and Meyer, 1980; Bichteler, 1981), that we can prove our main theorem, which we restate for convenience.

Theorem 5.29. *Let X be a càdlàg semimartingale, and $V \in \mathcal{L}$. Then, there exists a sequence $(V^n)_{n\in\mathbb{N}}$ in $\mathcal{S}(\psi)$ that converges to V, and such that*

$$\int V_s^n \,\mathrm{d}X_s \to \int V_s \,\mathrm{d}X_s, \quad n \to \infty,$$

where convergence in both cases holds in the topology of ucp-convergence.

Proof. We know from Protter (2005, Theorems II.4.10 and II.4.11) that the linear map

$$\mathcal{S} \ni V \mapsto I(V,X) = \int V_s \,\mathrm{d}X_s \in \mathcal{D} \tag{5.13}$$

is continuous between $(\mathcal{S}, \mathcal{T}_{\mathrm{ucp}})$ and $(\mathcal{D}, \mathcal{T}_{\mathrm{ucp}})$, and that it can be extended to a continuous map between $(\mathcal{L}, \mathcal{T}_{\mathrm{ucp}})$ and $(\mathcal{D}, \mathcal{T}_{\mathrm{ucp}})$. Proposition 5.28 shows that $\mathcal{S}(\psi)$ is dense in $(\mathcal{S}, \mathcal{T}_{\mathrm{ucp}})$, which concludes the proof. □

Remark 5.30. Theorem 5.29 is our main result. If we interpret elements of $\mathcal{S}(\psi)$ as trading strategies that use neural networks which map as an input past market data $(\tau, Y_{t_0}^{\tau}, \ldots, Y_{t_m}^{\tau})$ to a random output φ that represents a trading decision, then Theorem 5.29 tells us that we can approximate any other strategy $V \in \mathcal{L}$ and the corresponding wealth process $\int V \,\mathrm{d}X$ when trading with respect to a semimartingale X arbitrarily well with respect to $\mathcal{T}_{\mathrm{ucp}}$, see also Remark 5.31 below. In this sense algorithmic strategies can be seen as universal approximators for a large class of stochastic (integral) processes which are relevant for mathematical finance. A version of Theorem 5.29 for predictable integrands ins provided below.

Remark 5.31. Note that convergence of a sequence with respect to $\mathcal{T}_{\mathrm{ucp}}$ implies the existence of a subsequence, such that $\mathbb{P}$-almost surely, sample paths converge uniformly over compact time intervals. We may therefore pass to subsequences in Theorem 5.29 to obtain $\mathbb{P}$-almost surely pathwise convergence uniformly over compact time intervals.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

This statement can be seen as a stochastic process version of the statement that neural networks can approximate continuous functions uniformly on compacts.

For $p \in [1, \infty)$, we denote by $\mathcal{H}_0^p \subset \mathcal{H}^p$ the set of all $M \in \mathcal{H}^p$ with $M_0 = 0$. For $M \in \mathcal{H}_0^p$, denote by $L^p(M)$ the set of predictable processes V with

$$\mathbb{E}\Big[\Big(\int_0^\infty V_s^2 \,\mathrm{d}[M]_s\Big)^{p/2}\Big] < \infty,$$

which we endow with the seminorm $V \mapsto \|V\|_{L^p(M)} \coloneqq \mathbb{E}[(V^2 \bullet [M])_\infty^{p/2}]^{1/p}$.

Theorem 5.32. *Let $p \in [1, \infty)$, and assume that there exists $\varepsilon > 0$ such that $M \in \mathcal{H}_0^{p+\varepsilon}$. Then the set $\mathcal{E}(\psi)$ of elementary algorithmic strategies is dense in $L^p(M)$ with respect to the topology induced by the seminorm $\|\cdot\|_{L^p(M)}$. Moreover, for every $U \in L^p(M)$, there exists a sequence $(U^n)_{n\in\mathbb{N}}$ of elementary algorithmic strategies, such that*

$$\lim_{n\to\infty} \int U_s^n \,\mathrm{d}M_s = \int U_s \,\mathrm{d}M_s,$$

where convergence holds in $\mathcal{H}^p$.

Proof. Let $V \in L^p(M)$ and $V^n = V\mathbb{1}_{|V|\le n}$ for $n \in \mathbb{N}$. Using the dominated convergence theorem, we have $V^n \to V$ in $L^p(M)$. Now let $V \in L^p(M)$ be bounded. Using the functional monotone class theorem (Protter, 2005, Theorem 1.2.8), V may be approximated in $L^p(M)$ by a sequence from $\mathcal{E}$, since $\mathcal{E}$ generates the predictable σ-algebra, i.e. the σ-algebra generated by all left-continuous and $\mathbb{F}$-adapted processes.

What remains to be shown is that $\mathcal{E}(\psi)$ is dense in $L^p(M) \cap \mathcal{E}$. Let $V \in L^p(M) \cap \mathcal{E}$ and $\varepsilon > 0$. By Proposition 5.28, there exists $U \in \mathcal{E}(\psi)$, such that $\|(V-U)_\infty^*\|_q < \varepsilon^{1/p}$, where $q = p(p+\varepsilon)/\varepsilon$. Using the monotonicity of $[M]$ and Hölder's inequality with $\tilde{p} = (p+\varepsilon)/\varepsilon$ and $\tilde{q} = (p+\varepsilon)/p$,

$$\mathbb{E}\Big[\Big(\int_0^\infty (V_s - U_s)^2 \,\mathrm{d}[M]_s\Big)^{p/2}\Big] \le \mathbb{E}[((V-U)_\infty^*)^p [M]_\infty^{p/2}] \le \|((V-U)_\infty^*)^p\|_{\tilde{p}} \|[M]_\infty^{p/2}\|_{\tilde{q}}.$$

An application of the Burkholder–Davis–Gundy (BDG) inequality (Cohen and Elliott, 2015, Theorem 11.5.5) and Doob's L^p-inequality (Cohen and Elliott, 2015, Theorem 5.1.3) yields

$$\|((V-U)_\infty^*)^p\|_{\tilde{p}} \|[M]_\infty^{p/2}\|_{\tilde{q}} \lesssim \|(V-U)_\infty^*\|_q^p \|M_\infty^*\|_{p+\varepsilon}^p \lesssim \varepsilon \|M\|_{\mathcal{H}^{p+\varepsilon}}^p,$$

which shows that $\mathcal{E}(\psi)$ is dense in $L^p(M) \cap \mathcal{E}$.

In order to show the last claim, note that for each $U \in L^p(M)$, we have $\int U \,\mathrm{d}M \in \mathcal{H}^p$ (Cohen and Elliott, 2015, Corollary 12.3.6). Given $U \in L^p(M)$, there exists by Theorem 5.32 a sequence $(U^n)_{n\in\mathbb{N}}$ from $\mathcal{E}(\psi)$ that converges to U with respect to

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$\|\cdot\|_{L^p(M)}$. Using the fact that $[U \bullet M] = U^2 \bullet [M]$ as well as the BDG inequality, we have

$$\Big\| \int U_s \, \mathrm{d}M_s - \int U_s^n \, \mathrm{d}M_s \Big\|_{\mathcal{H}^p} \lesssim \|U - U^n\|_{L^p(M)},$$

which concludes the proof. □

5.3.2. Deep mean-variance hedging

For $M \in \mathcal{H}_0^2$, let $\mathcal{K}(M) \subset \mathcal{H}^2$ denote the stable subspace generated by M. We know that $\mathcal{K}(M) = \{V \bullet M : V \in L^2(M)\}$ (Cohen and Elliott, 2015, Theorem 12.2.7), and that (by definition) $\mathcal{K}(M)$ is closed in $\mathcal{H}^2$, hence $(\mathcal{K}(M), \|\cdot\|_{\mathcal{H}^2})$ is a Banach space.

The proof of the following proposition is a straightforward consequence of the results from the previous subsection.

Proposition 5.33. *In the context of Theorem 5.32 applied to the case $p = 2$, the set*

$$\mathcal{K}_\psi(M) := \{V \bullet M : V \in \mathcal{E}(\psi)\}$$

is dense in $\mathcal{K}(M)$.

As an alternative to the proof of Theorem 5.32, it is remarkable that there is another proof, which closely follows the proof of Theorem 1 in Hornik (1991).

Alternative proof of Proposition 5.33. Let us assume to the contrary that $\mathcal{K}_\psi(M)$ is not dense in $\mathcal{K}(M)$. As a consequence of the Hahn–Banach theorem (Rudin, 1987, Theorem 5.19), there exists an $F \in (\mathcal{K}(M))^*$ such that $F \neq 0$ and $F(V \bullet M) = 0$ for each $V \in \mathcal{E}(\psi)$. Recall that $(\mathcal{H}^2)^* \simeq \mathcal{H}^2$, and that $\mathcal{K}(M)$ is a closed subset of $\mathcal{H}^2$. Let $\mathcal{N}_0$ denote the subspace of all $F \in (\mathcal{H}^2)^*$ that annihilate $\mathcal{K}(M)$, i.e. for which $F(V \bullet M) = 0$ for each $V \in L^2(M)$. The space $(\mathcal{K}(M))^*$ can then be identified with the quotient space $(\mathcal{H}^2)^*/\mathcal{N}_0$. Let $N \in (\mathcal{H}^2)^*/\mathcal{N}_0$ be such that $F(X) = \mathbb{E}[X_\infty N_\infty]$ for each $X \in \mathcal{K}(M)$.

If $V \in \mathcal{E}$, then

$$\begin{aligned}
0 \leq |F(V \bullet M)| = |\mathbb{E}[(V \bullet M)_\infty N_\infty]| &\leq \sum_{i=1}^n |\mathbb{E}[\varphi_i (M_{t_{i+1}} - M_{t_i}) N_\infty]| \\
&\leq \sum_{i=1}^n |\mathbb{E}[(\varphi_i - \varphi_i^m)(M_{t_{i+1}} - M_{t_i}) N_\infty]| + \sum_{i=1}^n |\underbrace{\mathbb{E}[\varphi_i^m (M_{t_{i+1}} - M_{t_i}) N_\infty]}_{=0 \text{ since } F=0 \text{ on } \mathcal{K}_\psi(M)}| \\
&\leq \sum_{i=1}^n \|\varphi_i - \varphi_i^m\|_q \|M_{t_{i+1}} - M_{t_i}\|_{2+\varepsilon} \|N_\infty\|_2 \to 0, \quad m \to \infty,
\end{aligned}$$

where for $i \in \{1, \dots, n\}$ the sequence $(\varphi_i^m)_{m \in \mathbb{N}}$ consists of elements from $\mathcal{NN}_{t_i}(\psi)$ that converge to φ_i in $L^q(\mathbb{P})$, where q is chosen such that $1/(2+\varepsilon) + 1/q = 1$. The last

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

inequality follows by first applying the Cauchy–Schwarz inequality to separate $\|N_\infty\|_{L^2(\mathbb{P})}$, and then Hölder's inequality.

Similarly as in the proof of Theorem 5.32, an application of the functional monotone class and dominated convergence theorems shows that $F(V \bullet M) = 0$ extends from $V \in \mathcal{E}$ to $V \in L^2(M)$, which yields a contradiction. □

Remark 5.34. Here, one could say that elementary algorithmic strategies are "discriminatory" (see Proposition 5.15), in the sense that there cannot exist $F \in (\mathcal{K}(M))^*$ with $F \neq 0$ and $F(V \bullet M) = 0$ for every $V \in \mathcal{E}(\psi)$.

Theorem 5.35 (Deep mean-variance hedging)**.** *In the context of Theorem 5.32 applied to the case $p = 2$, let $H \in L^2(\mathbb{P})$ be a real-valued $\mathcal{F}_\infty$-measurable claim, and assume $\mathcal{F}_0$ to be trivial. Then:*

$$\min_{c\in\mathbb{R},\, U\in L^2(M)} \mathbb{E}\left[\left(H - c - \int_0^\infty U_t \,\mathrm{d}M_t\right)^2\right] = \inf_{c\in\mathbb{R},\, U\in\mathcal{E}(\psi)} \mathbb{E}\left[\left(H - c - \int_0^\infty U_t \,\mathrm{d}M_t\right)^2\right]. \quad (5.14)$$

More precisely, if $U^* \in L^2(M)$ denotes the mean-variance hedging strategy that minimizes the quadratic hedging error on the left-hand side of Eq. (5.14), and $c^* \in \mathbb{R}$ is the optimal initial capital, then, for each $\delta > 0$, there exists $U \in \mathcal{E}(\psi)$, such that

$$\mathbb{E}\left[\left(H - c^* - \int_0^\infty U_t \,\mathrm{d}M_t\right)^2\right] < \mathbb{E}\left[\left(H - c^* - \int_0^\infty U_t^* \,\mathrm{d}M_t\right)^2\right] + \delta.$$

Proof. Let $V \in \mathcal{H}^2$ be the martingale that is closable by H, i.e. $V_t = \mathbb{E}[H \,|\, \mathcal{F}_t]$ for every $t \in \mathbb{R}_+$. It follows from the Kunita–Watanabe decomposition, that there exists a unique $U^* \in \mathcal{K}(M)$ such that, denoting $c^* = \mathbb{E}[H]$, V admits the orthogonal decomposition $V = c^* + U^* \bullet M + N$, where $N \in (\mathcal{K}(M))^\perp$, and such that c^* and U^* minimize the quadratic hedging error (Rheinländer and Sexton, 2011, Theorem 6.2).

By Proposition 5.33, we can find $U \in \mathcal{E}(\psi)$ such that $\|U^* \bullet M - U \bullet M\|_{\mathcal{H}^2} < \delta$, hence

$$\begin{aligned}\|H - c^* - (U \bullet M)_\infty\|_2 &\leq \|H - c^* - (U^* \bullet M)_\infty\|_2 + \|(U^* \bullet M)_\infty - (U \bullet M)_\infty\|_2 \\ &\leq \|H - c^* - (U^* \bullet M)_\infty\|_2 + \|U^* \bullet M - U \bullet M\|_{\mathcal{H}^2} \\ &< \|H - c^* - (U^* \bullet M)_\infty\|_2 + \delta,\end{aligned}$$

which concludes the proof. □

5.3.3. No free lunch with vanishing risk for algorithmic strategies

We recall Théorème 3 from Stricker (1990). We restrict all processes to the time-index set $[0, 1]$, fix $p \in [1, \infty)$, and let q be such that $1/p + 1/q = 1$ (if $p = 1$ then take $q = \infty$). Let $X = (X_t)_{t\in[0,1]}$ denote a continuous, $\mathbb{R}^d$-valued and adapted process with $X_t \in L^p(\mathbb{P})$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

for every $t \in [0,1]$. Note that in this case, the set $\mathcal{K}_{\mathrm{e}}(X) = \{V \bullet X \colon V \in \mathcal{E}\}$ consists of well-defined processes.

Theorem 5.36. *There exists an equivalent probability measure $\mathbb{Q} \sim \mathbb{P}$ with $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} \in L^q(\mathbb{P})$ such that X is an $(\mathbb{F},\mathbb{Q})$-martingale if and only if $\overline{\mathcal{K}_{\mathrm{e}}(X)} \cap L^p_+(\mathbb{P}) = \{0\}$, where the closure $\overline{\mathcal{K}_{\mathrm{e}}(X)}$ is taken in $L^p(\mathbb{P})$, and $L^p_+(\mathbb{P}) = \{f \in L^p(\mathbb{P}) \colon f \geq 0\}$.*

The condition $\overline{\mathcal{K}_{\mathrm{e}}(X)} \cap L^p_+(\mathbb{P}) = \{0\}$ may be referred to as no free lunch with vanishing risk (NFLVR) for elementary integrands, see also Delbaen and Schachermayer (1994). Combining this theorem with our results gives rise to a NFLVR condition for algorithmic strategies.

Theorem 5.37. *Assume that there exists $\varepsilon > 0$ such that $X_t \in L^{p+\varepsilon}(\mathbb{P})$ for every $t \in [0,1]$. Then there exists an equivalent probability measure $\mathbb{Q} \sim \mathbb{P}$ with $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} \in L^q(\mathbb{P})$ such that X is an $(\mathbb{F},\mathbb{Q})$-martingale if and only if $\overline{\mathcal{K}_{\psi}(X)} \cap L^p_+(\mathbb{P}) = \{0\}$.*

Proof. We just need to show that $\mathcal{K}_{\mathrm{e}}(X) \subset \overline{\mathcal{K}_{\psi}(X)}$. Let $V \in \mathcal{E}$ have a representation $V = \sum_{i=1}^n \varphi_i \mathbb{1}_{(t_i,t_{i+1}]}$, such that $V \bullet M = \sum_{i=1}^n \varphi_i (M_{t_{i+1}} - M_{t_i})$. Let $r > 1$ be sufficiently large and, for every $i \in \{1,\dots,n\}$, let $(\varphi_{i,m})_{m\in\mathbb{N}}$ be a sequence from $\mathcal{NN}_{t_i}(\psi)$ that converges to φ_i in $L^r(\mathbb{P})$. Let $V^m = \sum_{i=1}^n \varphi_{i,m} \mathbb{1}_{(t_i,t_{i+1}]}$. An application of Hölder's inequality shows that

$$\|V \bullet X - V^m \bullet X\|_p \leq \sum_{i=1}^n \|\varphi_i - \varphi_{i,m}\|_r \|X_{t_{i+1}} - X_{t_i}\|_{p+\varepsilon},$$

where the right-hand side converges to zero as $m \to \infty$. □

Conclusions

In this chapter, we have established the fundamental role of algorithmic trading strategies in continuous-time hedging and stochastic integration. By interpreting algorithmic strategies as simple predictable step processes, we derived a universal approximation theorem for stochastic (integral) processes. We also demonstrated that algorithmic strategies can be used for continuous-time mean-variance hedging, and presented a no free lunch with vanishing risk condition. The theoretical framework developed here provides a foundation for future explorations into the applications of machine learning in financial mathematics from a theoretical and continuous-time point of view.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Conclusions and Outlook on Future Research Directions

This thesis explored connections between machine learning and financial and actuarial mathematics, and showed that there is ample scope for fascinating use cases. The results are categorized into three parts covering both quantitative finance and actuarial science, as well as theory and practical applications. The first part demonstrated how deep learning can help to find optimal measures, here exemplified through the concept of importance sampling. In the second part we used the approximation capabilities of neural networks to tackle complex input-output maps in the context of a high-dimensional stochastic control problem. Finally, the third part demonstrated the ability of neural networks to process information in order to make informed decisions, and we used this to study a novel reinsurance problem as well as stochastic integration in continuous-time.

Besides importance sampling, we have seen that there is potential for applying neural networks in the context of stratified sampling. Moreover, the results presented in Chapter 1 suggest an extension to the field of McKean–Vlasov equations. Another idea would be the application of the concept of deep measure projections to the field of arbitrage-free pricing in incomplete financial markets.

The results presented in Chapter 3 demonstrate that the deep LSMC method is useful for solving complex, high-dimensional stochastic control problems. A follow-up to Chapter 3 could see an extension to a stochastic version of the regional variant of the DICE model, the RICE model. Insights into the impact of uncertainty on climate-economy models could become useful for performing actuarial or credit-risk stress testing. Neural networks could also be used to approximate the pricing map for some liquid products on the life insurance and annuity market, which could allow an extraction of market-implied mortality rates.

The idea of using algorithmic strategies to study optimal reinsurance could be extended to the optimal choice of premium rates subject to a premium-dependent insurance demand. Moreover, Chapter 5 has seen deep mean-variance hedging as an application, and the results from Chapter 5 suggest a further extension to hedging in Orlicz spaces. Finally the continuous-time nature of Chapter 5 motivates the joint study of optimal algorithmic trades and optimal trading times, thereby combining the concepts of deep hedging and deep optimal stopping.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Bibliography

Ackerman, F., DeCanio, S. J., Howarth, R. B., and Sheeran, K. Limitations of integrated assessment models of climate change. *Climatic Change*, 95(3):297–315, 2009. 81

Aïd, R., Campi, L., Langrené, N., and Pham, H. A probabilistic numerical method for optimal multiple switching problems in high dimension. *SIAM Journal on Financial Mathematics*, 5(1):191–231, 2014. 82

Albrecher, H. and Thonhauser, S. Optimality results for dividend problems in insurance. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas*, 103(2):295–320, 2009. 110

Albrecher, H., Beirlant, H., and Teugels, J. L. *Reinsurance: Actuarial and Statistical Aspects*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2017. 110

Andréasson, J. G. and Shevchenko, P. V. A bias-corrected least-squares Monte Carlo for solving multi-period utility models. *European Actuarial Journal*, 12(1):349–379, 2022. 82, 90, 93

Andréasson, J. G. and Shevchenko, P. V. Optimal annuitisation, housing and reverse mortgage in retirement in the presence of a means-tested public pension. *European Actuarial Journal*, 14:871–904, 2024. 82

Andresen, A., Imkeller, P., and Perkowski, N. Large deviations for Hilbert-space-valued Wiener processes: A sequence space approach. In *Malliavin Calculus and Stochastic Analysis*, volume 34 of *Springer Proceedings in Mathematics & Statistics*, pages 115–138. Springer, 2013. 24

Ansel, J.-P. and Stricker, C. Décomposition de Kunita–Watanabe. In *Séminaire de probabilités de Strasbourg*, volume 27, pages 30–32. Springer, 1993. 125

Arandjelović, A., Kingston, G., and Shevchenko, P. V. Life cycle insurance, bequest motives and annuity loads. *Journal of Economic Dynamics and Control*, 157, 2023. iv

Arandjelović, A., Rheinländer, T., and Shevchenko, P. V. Importance sampling for option pricing with feedforward neural networks. *Finance and Stochastics*, 29:97–141, 2025. iv, 129

Arnold, V. I. On functions of three variables. *Doklady Akademii Nauk SSSR*, 114:679–681, 1957. 1

Arouna, B. Robbins–Monro algorithms and variance reduction in finance. *Journal of Computational Finance*, 7(2):35–62, 2003. 31, 36

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Asmussen, S. and Albrecher, H. *Ruin Probabilities*, volume 14 of *Advanced Series on Statistical Science & Applied Probability*. World Scientific Publishing, 2nd edition, 2010. 110

Avanzi, B. Strategies for dividend distribution: A review. *North American Actuarial Journal*, 13(2):217–251, 2009. 110

Bach, F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. 128

Bai, L. and Guo, J. Optimal proportional reinsurance and investment with multiple risky assets and no-shorting constraint. *Insurance: Mathematics and Economics*, 42 (3):968–975, 2008. 110

Baldi, P., Ben Arous, G., and Kerkyacharian, G. Large deviations and the Strassen theorem in Hölder norm. *Stochastic Processes and their Applications*, 42(1):171–180, 1992. 24

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. 115

Becker, S., Cheridito, P., and Jentzen, A. Deep optimal stopping. *Journal of Machine Learning Research*, 20(74):1–25, 2019. 116

Belomestny, D. Pricing Bermudan options by nonparametric regression: Optimal rates of convergence for lower estimates. *Finance and Stochastics*, 15:655–683, 2011. 82

Belomestny, D., Kolodko, A., and Schoenmakers, J. Regression methods for stochastic control problems and their convergence analysis. *SIAM Journal on Control and Optimization*, 48(5):3562–3588, 2010. 82

Bichteler, K. Stochastic integrators. *Bulletin (New Series) of the American Mathematical Society*, 1(5):761–765, 1979. 144

Bichteler, K. Stochastic integration and L^p-theory of semimartingales. *The Annals of Probability*, 9(1):49–89, 1981. 144

Billingsley, P. *Probability and Measure*. Wiley Series in Probability and Statistics. John Wiley & Sons, anniversary edition, 2012. 45, 128, 130

Bogachev, V. I. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, 1998. 19, 42, 43, 44, 49

Bogachev, V. I. *Measure Theory, Volume 2*. Springer, 1st edition, 2007. 127, 128, 134

Borch, K. The utility concept applied to the theory of insurance. *ASTIN Bulletin: The Journal of the IAA*, 1(5):245–255, 1961. 110

Braunsteins, P. and Mandjes, M. The Cramér–Lundberg model with a fluctuating number of clients. *Insurance: Mathematics and Economics*, 112:1–22, 2023. 109

Brown, J. R. and Finkelstein, A. Why is the market for long-term care insurance so small? *Journal of Public Economics*, 91(10):1967–1991, 2007. 62, 63, 65

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Buehler, H., Gonon, L., Teichmann, J., and Wood, B. Deep hedging. *Quantitative Finance*, 19(8):1271–1291, 2019. 5, 116, 117, 124, 125, 126, 138

Cai, Y. and Lontzek, T. S. The social cost of carbon with economic and climate risks. *Journal of Political Economy*, 127(6):2684–2734, 2019. 81, 82, 83, 96, 98

Cameron, R. H. and Martin, W. T. Transformations of Wiener integrals under translations. *Annals of Mathematics*, 45(2):386–396, 1944. 44

Capriotti, L. Least-squares importance sampling for Monte Carlo security pricing. *Quantitative Finance*, 8(5):485–497, 2008. 31, 36, 38

Carroll, C. D. Why do the rich save so much? Working Paper 6549, National Bureau of Economic Research, 1998. 58

Cauchy, A.-L. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus hebdomadaires des séances de l'Académie des sciences*, 25:536–538, 1847. 1

Ceci, C., Colaneri, K., and Cretarola, A. Optimal reinsurance and investment under common shock dependence between financial and actuarial markets. *Insurance: Mathematics and Economics*, 105:252–278, 2022. 109

Cheng, X., Jin, Z., and Yang, H. Optimal insurance strategies: A hybrid deep learning Markov chain approximation approach. *ASTIN Bulletin: The Journal of the IAA*, 50 (2):449–477, 2020. 111

Cheridito, P. and Li, T. Risk measures on Orlicz hearts. *Mathematical Finance*, 19(2): 189–214, 2009. 136

Cherny, A. and Shiryaev, A. N. Vector stochastic integrals and the fundamental theorem of asset pricing. *Proceedings of the Steklov Mathematical Institute*, 237:12–56, 2002. 14, 15, 45, 46, 50

Cheung, E. C. K. and Liu, H. Joint moments of discounted claims and discounted perturbation until ruin in the compound Poisson risk model with diffusion. *Probability in the Engineering and Informational Sciences*, 37(2):387–417, 2023. 109

Ciesielski, Z. On the isomorphisms of the spaces H_α and m. *Bulletin de l'Académie Polonaise des Sciences. Série des Sciences Mathématiques, Astronomiques et Physiques*, 8:217–222, 1960. 24

Ciesielski, Z., Kerkyacharian, G., and Roynette, B. Quelques espaces fonctionnels associés à des processus gaussiens. *Studia Mathematica*, 107(2):171–204, 1993. 25

Clément, E., Lamberton, D., and Protter, P. E. An analysis of a least squares regression method for American option pricing. *Finance and Stochastics*, 6(4):449–471, 2002. 42

Cohen, S. N. and Elliott, R. J. *Stochastic Calculus and Applications*. Probability and Its Applications. Birkhäuser, 2nd edition, 2015. 13, 16, 25, 27, 51, 124, 130, 138, 142, 145, 146

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989. 1, 20, 23, 51, 93, 116, 124, 134

Davies, P. I. and Higham, N. J. Numerically stable generation of correlation matrices and their factors. *BIT Numerical Mathematics*, 40:640–651, 2000. 40

Davis, M. H. and Norman, A. Portfolio selection with transaction costs. *Mathematics of Operations Research*, 15(4):676–713, 1990. 57

De Nardi, M., French, E., and Jones, J. B. Why do the elderly save? The role of medical expenses. *Journal of Political Economy*, 118(1):39–75, 2010. 58

Delbaen, F. Representing martingale measures when asset prices are continuous and bounded. *Mathematical Finance*, 2(2):107–130, 1992. 126

Delbaen, F. and Schachermayer, W. A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, 300:463–520, 1994. 126, 148

Dellacherie, C. and Meyer, P.-A. *Probabilités et potentiel. Chapitres V à VIII.* Hermann, 1980. 144

Der, R. and Lee, D. Beyond Gaussian processes: On the distributions of infinite networks. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18, 2005. 128

dos Reis, G., Smith, G., and Tankov, P. Importance sampling for McKean–Vlasov SDEs. *Applied Mathematics and Computation*, 453, 2023. 10, 42

Dufresne, F. and Gerber, H. U. Risk theory for the compound Poisson process that is perturbed by diffusion. *Insurance: Mathematics and Economics*, 10(1):51–59, 1991. 109

Dvoretzky, A. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 39–55. University of California Press, 1956. 94, 116

Eisenberg, J. Optimal dividends under a stochastic interest rate. *Insurance: Mathematics and Economics*, 65:259–266, 2015. 109

Elstrodt, J. *Maß- und Integrationstheorie*. Springer, 8th edition, 2018. 51

Feigenbaum, J., Gahramanov, E., and Tang, X. Is it really good to annuitize? *Journal of Economic Behavior & Organization*, 93:116–140, 2013. 62

Feinstein, Z. and Rudloff, B. Deep learning the efficient frontier of convex vector optimization problems. *Journal of Global Optimization*, 90:429–458, 2024. 113

Filipović, D. *Consistency Problems for Heath–Jarrow–Morton Interest Rate Models*. Lecture Notes in Mathematics. Springer, 2001. 18

Fischer, S. A life cycle model of life insurance purchases. *International Economic Review*, 14(1):132–152, 1973. 56, 67, 68

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Fleming, W. H. and Soner, H. M. *Controlled Markov Processes and Viscosity Solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2006. 78, 91

Friedl, A., Kübler, F., Scheidegger, S., and Usui, T. Deep uncertainty quantification: With an application to integrated assessment models. Technical report, Working Paper University of Lausanne, 2023. 97

Friz, P. K. and Hairer, M. *A Course on Rough Paths*. Universitext. Springer, 2nd edition, 2020. 24

Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989. 20, 124

Genin, A. and Tankov, P. Optimal importance sampling for Lévy processes. *Stochastic Processes and their Applications*, 130(1):20–46, 2020. 42

Gerber, H. U. An extension of the renewal equation and its application in the collective theory of risk. *Scandinavian Actuarial Journal*, 1970(3–4):205–210, 1970. 109

Gerber, H. U. and Shiu, E. S. On the time value of ruin. *North American Actuarial Journal*, 2(1):48–72, 1998. 110, 115

Gillingham, K., Nordhaus, W. D., Anthoff, D., Blanford, G., Bosetti, V., Christensen, P., McJeon, H., Reilly, J., and Sztorc, P. Modeling uncertainty in climate change: A multi-model comparison. Working Paper 21637, National Bureau of Economic Research, 2015. 83

Glasserman, P. *Monte Carlo Methods in Financial Engineering*, volume 53 of *Stochastic Modelling and Applied Probability*. Springer, 2003. 30, 33

Glasserman, P. and Wang, Y. Counterexamples in importance sampling for large deviations probabilities. *The Annals of Applied Probability*, 7(3):731–746, 1997. 36

Glasserman, P., Heidelberger, P., and Shahabuddin, P. Asymptotically optimal importance sampling and stratification for pricing path-dependent options. *Mathematical Finance*, 9(2):117–152, 1999. 10, 31, 33, 37, 38

Gowurin, M. Über die Stieltjessche Integration abstrakter Funktionen. *Fundamenta Mathematicae*, 27(1):254–265, 1936. 47

Grubb, M., Wieners, C., and Yang, P. Modeling myths: On DICE and dynamic realism in integrated assessment models of climate change mitigation. *Wiley Interdisciplinary Reviews: Climate Change*, 12(3), 2021. 81

Guasoni, P. and Robertson, S. Optimal importance sampling with explicit formulas in continuous time. *Finance and Stochastics*, 12(1):1–19, 2008. 10, 26, 27, 28, 31, 33, 34, 41

Gühring, I., Kutyniok, G., and Petersen, P. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859, 2020. 25

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Hakansson, N. H. Optimal investment and consumption strategies under risk, an uncertain lifetime, and insurance. *International Economic Review*, 10(3):443–466, 1969. 56

He, S.-w., Wang, J.-g., and Yan, J.-a. *Semimartingale Theory and Stochastic Calculus*. CRC Press, 1992. 124, 139

Hecht-Nielsen, R. Theory of the backpropagation neural network. In *International 1989 Joint Conference on Neural Networks*, pages 593–605, 1989. 124

Heijdra, B. J., Mierau, J. O., and Reijnders, L. S. A tragedy of annuitization? Longevity insurance in general equilibrium. *Macroeconomic Dynamics*, 18(7):1607–1634, 2014. 62

Hipp, C. Company value with ruin constraint in Lundberg models. *Risks*, 6(3):73, 2018. 110

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. 1, 116

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. 1, 20, 22, 23, 51, 93, 94, 116, 118, 124, 126, 134, 135, 146

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. 20, 124

Horvath, B., Muguruza, A., and Tomas, M. Deep learning volatility: A deep neural network perspective on pricing and calibration in (rough) volatility models. *Quantitative Finance*, 21(1):11–27, 2021. 116

Huang, G.-B., Chen, L., and Siew, C.-K. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks*, 17(4):879–892, 2006. 116

Huang, H. and Milevsky, M. A. Portfolio choice and mortality-contingent claims: The general HARA case. *Journal of Banking and Finance*, 32(11):2444–2452, 2008. 61

Human Mortality Database. Data downloaded on 13 December 2022. Institutional sponsors: Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France), 2022. 73, 74

Ikefuji, M., Laeven, R. J. A., Magnus, J. R., and Muris, C. Expected utility and catastrophic risk in a stochastic economy-climate model. *Journal of Econometrics*, 214 (1):110–129, 2020. 83, 96

Jacod, J. *Calcul Stochastique et Problèmes de Martingales*, volume 714 of *Lecture Notes in Mathematics*. Springer, 1979. 14, 15, 16, 45, 129

Jacod, J. Intégrales stochastiques par rapport à une semi-martingale vectorielle et changements de filtration. In *Séminaire de probabilités de Strasbourg*, volume 14, pages 161–172. Springer, 1980. 14

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Jensen, S. and Traeger, C. P. Optimal climate change mitigation under long-term growth uncertainty: Stochastic integrated assessment and analytic findings. *European Economic Review*, 69:104–125, 2014. 96, 99

Jin, Z., Yang, H., and Yin, G. A hybrid deep learning method for optimal insurance strategies: Algorithms and convergence analysis. *Insurance: Mathematics and Economics*, 96:262–275, 2021. 111

Jourdain, B. and Lelong, J. Robust adaptive importance sampling for normal random vectors. *The Annals of Applied Probability*, 19(5):1687–1718, 2009. 39, 40

Kallenberg, O. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, 3rd edition, 2021. 13, 45, 51, 125, 131, 132, 138, 141

Karatzas, I. and Shreve, S. E. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer, 2 edition, 1998. 139

Kawai, R. Optimal importance sampling parameter search for Lévy processes via stochastic approximation. *SIAM Journal on Numerical Analysis*, 47(1):293–307, 2009. 42

Kazamaki, N. *Continuous Exponential Martingales and BMO*. Lecture Notes in Mathematics. Springer, 1994. 130

Kelly, D. L. and Kolstad, C. D. Bayesian learning, growth, and pollution. *Journal of Economic Dynamics and Control*, 23(4):491–518, 1999. 81, 83, 87

Kelly, D. L. and Kolstad, C. D. Solving infinite horizon growth models with an environmental sector. *Computational Economics*, 18:217–231, 2001. 81, 99

Kharroubi, I., Langrené, N., and Pham, H. A numerical algorithm for fully nonlinear HJB equations: An approach by control randomization. *Monte Carlo Methods and Applications*, 20(2):145–165, 2014. 82, 90

Kidger, P. and Lyons, T. Universal approximation with deep narrow networks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327, 2020. 20, 21, 22, 116

Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952. 94, 116

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. Technical report, arXiv preprint, https://arxiv.org/abs/1412.6980, cs.LG, 2017. 32, 98, 119

Kolmogorov, A. N. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Doklady Akademii Nauk SSSR*, 108:179–182, 1956. 1

Kolmogorov, A. N. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk SSSR*, 114:953–956, 1957. 1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Kratsios, A. The universal approximation property. *Annals of Mathematics and Artificial Intelligence*, 89:435–469, 2021. 20, 93, 116, 124

Kratsios, A. and Bilokopytov, I. Non-Euclidean universal approximation. In *Advances in Neural Information Processing Systems*, volume 33, pages 10635–10646. Curran Associates, 2020. 119

Krickeberg, K. Convergence of conditional expectation operators. *Theory of Probability & Its Applications*, 9(4):538–549, 1964. 133

Kunita, H. and Watanabe, S. On square integrable martingales. *Nagoya Mathematical Journal*, 30:209–245, 1967. 125

Leach, A. J. The climate change learning curve. *Journal of Economic Dynamics and Control*, 31(5):1728–1752, 2007. 81, 83

Leibniz, G. W. Calculus tangentium differentialis, November 1676. 1

Leimcke, G. *Bayesian Optimal Investment and Reinsurance to Maximize Exponential Utility of Terminal Wealth*. PhD thesis, Karlsruher Institut für Technologie, 2020. 109

Lemaire, V. and Pagès, G. Unconstrained recursive importance sampling. *The Annals of Applied Probability*, 20(3):1029–1067, 2010. 11, 29, 30, 31, 53, 54

Lenart, A. The moments of the Gompertz distribution and maximum likelihood estimation of its parameters. *Scandinavian Actuarial Journal*, 2014(3):255–277, 2014. 74, 75

Lépingle, D. and Mémin, J. Intégrabilité uniforme dans L^r des martingales exponentielles. *Publications des séminaires de mathématiques et informatique de Rennes*, 1(9), 1978. 27

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. 20, 22, 124

Li, L., Mendoza-Arriaga, R., Mo, Z., and Mitchell, D. Modelling electricity prices: A time change approach. *Quantitative Finance*, 16(7):1089–1109, 2016. 34

Liao, Y., Fang, S.-C., and Nuttle, H. L. Relaxed conditions for radial-basis function networks to be universal approximators. *Neural Networks*, 16(7):1019–1028, 2003. 20

Lifshits, M. *Lectures on Gaussian Processes*. SpringerBriefs in Mathematics. Springer, 2012. 19, 42, 44, 47

Linnainmaa, S. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's thesis, University of Helsinki, 1970. 1

Lockwood, L. M. Bequest motives and the annuity puzzle. *Review of Economic Dynamics*, 15(2):226–243, 2012. 58, 61, 67

Lockwood, L. M. Incidental bequests and the choice to self-insure late-life risks. *American Economic Review*, 108(9):2513–2550, 2018. 58, 60, 61, 63, 65, 67, 72

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Longstaff, F. A. and Schwartz, E. S. Valuing American options by simulation: A simple least-squares approach. *The Review of Financial Studies*, 14(1):113–147, 2001. 4, 42, 82

Lontzek, T. S., Cai, Y., Judd, K. L., and Lenton, T. M. Stochastic integrated assessment of climate tipping points indicates the need for strict climate policy. *Nature Climate Change*, 5(5):441–444, 2015. 83

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6232–6240, 2017. 23

Lundberg, F. *I. Approximerad framställning af sannolikhetsfunktionen. II. Aterförsäkring af Kollektivrisker*. PhD thesis, 1903. 109

Magill, M. J. and Constantinides, G. M. Portfolio selection with transactions costs. *Journal of Economic Theory*, 13(2):245–263, 1976. 57

McCulloch, W. S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943. 1

Memin, J. Espaces de semi martingales et changement de probabilité. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 52(1):9–39, 1980. 14, 15

Mercer, R. Dense G_δ's contain orthonormal bases. *Proceedings of the American Mathematical Society*, 97(3):449–452, 1986. 49

Merton, R. C. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 3(4):373–413, 1971. 56

Mhaskar, H. N. and Micchelli, C. A. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, 13(3):350–373, 1992. 20

Milgram, M. S. The generalized integro-exponential function. *Mathematics of Computation*, 44(170):443–458, 1985. 63

Mincer, J. A. *Schooling, Experience, and Earnings*. National Bureau of Economic Research, 1974. 58, 76

Müller, T., McWilliams, B., Rousselle, F., Gross, M., and Novák, J. Neural importance sampling. *ACM Transactions on Graphics*, 38(5):1–19, 2019. 11

Murphy, K. M. and Welch, F. Empirical age-earnings profiles. *Journal of Labor Economics*, 8(2):202–229, 1990. 77

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer, 1996. 128

Nguyen, X., Wainwright, M. J., and Jordan, M. I. On surrogate loss functions and f-divergences. *The Annals of Statistics*, 37(2):876–904, 2009. 115

Nordhaus, W. D. The 'DICE' model: Background and structure of a dynamic integrated climate-economy model of the economics of global warming. Technical report, Cowles Foundation for Research in Economics at Yale University, 1992. 81

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Nordhaus, W. D. Projections and uncertainties about climate change in an era of minimal climate policies. *American Economic Journal: Economic Policy*, 10(3):333–360, 2018. 82, 83, 84, 87, 88, 90, 96, 98, 106

Nordhaus, W. D. and Yang, Z. A regional dynamic general-equilibrium model of alternative climate-change strategies. *The American Economic Review*, 86(4):741–765, 1996. 107

Øksendal, B. *Stochastic Differential Equations*. Universitext. Springer, 6th edition, 2003. 140

on Social Cost of Greenhouse Gases, I. W. G. Technical support document: Social cost of carbon for regulatory impact analysis under executive order 12866. Technical report, United States Government, 2016. 81

Park, J. and Sandberg, I. W. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257, 1991. 20

Park, J. and Sandberg, I. W. Approximation and radial-basis-function networks. *Neural Computation*, 5(2):305–316, 1993. 21

Pashchenko, S. Accounting for non-annuitization. *Journal of Public Economics*, 98:53–67, 2013. 62

Peijnenburg, K., Nijman, T., and Werker, B. J. The annuity puzzle remains a puzzle. *Journal of Economic Dynamics and Control*, 70:18–35, 2016. 62

Pindyck, R. S. The use and misuse of models for climate policy. *Review of Environmental Economics and Policy*, 11(1):100–114, 2017. 81, 107

Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999. 21

Pliska, S. R. and Ye, J. Optimal life insurance purchase and consumption/investment under uncertain lifetime. *Journal of Banking and Finance*, 31(5):1307–1319, 2007. 59, 61, 71, 72

Promislow, S. D. and Young, V. R. Minimizing the probability of ruin when claims follow Browian motion with drift. *North American Actuarial Journal*, 9(3):110–128, 2005. 110

Protter, P. E. *Stochastic Integration and Differential Equations*. Stochastic Modelling and Applied Probability. Springer, 2nd edition, 2005. 124, 144, 145

Rao, M. M. and Ren, Z. D. *Theory of Orlicz Spaces*. Marcel Dekker, 1991. 131, 132, 134, 136, 143

Reid, M. D. and Williamson, R. C. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010. 115

Rheinländer, T. Stochastic exponential. In *Encyclopedia of Quantitative Finance*. John Wiley & Sons, 2010. 27

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Rheinländer, T. and Schweizer, M. On L^2-projections on a space of stochastic integrals. *The Annals of Probability*, 25(4):1810–1831, 1997. 125

Rheinländer, T. and Sexton, J. *Hedging Derivatives*, volume 15 of *Advanced Series on Statistical Science & Applied Probability*. World Scientific, 2011. 147

Richard, S. F. Optimal consumption, portfolio and life insurance rules for an uncertain lived individual in a continuous time model. *Journal of Financial Economics*, 2(2): 187–203, 1975. 56, 78

Richman, R. AI in actuarial science - A review of recent advances - Part 2. *Annals of Actuarial Science*, 15(2):230–258, 2021. 110

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. 30, 32, 94, 116

Robertson, S. Sample path large deviations and optimal importance sampling for stochastic volatility models. *Stochastic Processes and their Applications*, 120(1):66–83, 2010. 10, 11, 26, 41

Ronen, E. and Ohad, S. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 907–940, 2016. 23

Rudin, W. *Real and Complex Analysis*. McGraw-Hill, 3rd edition, 1987. 136, 146

Rudin, W. *Functional Analysis*. McGraw-Hill, 2nd edition, 1991. 127

Rumerlhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. 1, 94, 116

Salem, R. On some singular monotonic functions which are strictly increasing. *Transactions of the American Mathematical Society*, 53(3):427–439, 1943. 15

Schilling, R. L. *Measures, Integrals and Martingales*. Cambridge University Press, 2nd edition, 2017. 132, 133

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, 61: 85–117, 2015. 124

Schmidli, H. Optimal proportional reinsurance policies in a dynamic setting. *Scandinavian Actuarial Journal*, 2001(1):55–68, 2001. 110, 119, 120

Schmidli, H. On minimizing the ruin probability by investment and reinsurance. *The Annals of Applied Probability*, 12(3):890–907, 2002. 110

Schmidli, H. *Stochastic Control in Insurance*. Probability and Its Applications. Springer, 2008. 110, 120

Schmock, U. Lecture Notes in Stochastic Analysis for Financial and Actuarial Mathematics. TU Wien, October 2024. 13, 24, 26, 45, 51, 128, 129, 130, 132, 133, 135, 141, 143

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Shevchenko, P. V., Murakami, D., Matsui, T., and Myrvoll, T. A. Impact of COVID-19 type events on the economy and climate under the stochastic DICE model. *Environmental Economics and Policy Studies*, 24:459–476, 2022. 83

Shiryaev, A. N. *Optimal Stopping Rules*, volume 8 of *Stochastic Modelling and Applied Probability*. Springer, 2008. 139

Shreve, S. E. and Soner, H. M. Optimal investment and consumption with transaction costs. *The Annals of Applied Probability*, 4(3):609–692, 1994. 57

Singer, I. Linear functionals on the space of continuous mappings of a compact Hausdorff space into a Banach space. *Revue de Mathématiques Pure et Appliquées*, 2:301–315, 1957. 47

Sobol', I. M. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. 100

Sobol', I. M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1–3):271–280, 2001. 95

Stricker, C. Arbitrage et lois de martingale. *Annales de l'I.H.P. Probabilités et statistiques*, 26(3):451–460, 1990. 6, 126, 147

Stroock, D. W. *Probability Theory*. Cambridge University Press, 2nd edition, 2011. 19, 42

Stroock, D. W. and Varadhan, S. *Multidimensional Diffusion Processes*. Classics in Mathematics. Springer, 2006. 139

Su, Y. and Fu, M. C. Importance sampling in derivative securities pricing. In *2000 Winter Simulation Conference Proceedings*, volume 1, pages 587–596, 2000. 31, 36

Taksar, M. I. and Markussen, C. Optimal dynamic reinsurance policies for large insurance portfolios. *Finance and Stochastics*, 7:97–121, 2003. 110

Thonhauser, S. and Albrecher, H. Dividend maximization under consideration of the time value of ruin. *Insurance: Mathematics and Economics*, 41(1):163–184, 2007. 110

Traeger, C. P. A 4-stated DICE: Quantitatively addressing uncertainty effects in climate change. *Environmental and Resource Economics*, 59(1):1–37, 2014. 81, 83, 99

Tsai, C. C.-L. On the discounted distribution functions of the surplus process perturbed by diffusion. *Insurance: Mathematics and Economics*, 28(3):401–419, 2001. 109

Tsitsiklis, J. N. and Van Roy, B. Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, 12(4):694–703, 2001. 82

Unser, M. Ridges, neural networks, and the Radon transform. *Journal of Machine Learning Research*, 24(37):1–33, 2023. 128

Weitzman, M. L. Fat-tailed uncertainty in the economics of catastrophic climate change. *Review of Environmental Economics and Policy*, 5(2):275–292, 2011. 81

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Werbos, P. J. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974. 1

Wiener, N. *Cybernetics or Control and Communication in the Animal and the Machine*. The MIT Press, 2nd edition, 1948. 1

Wiese, M., Knobloch, R., Korn, R., and Kretschmer, P. Quant GANs: Deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020. 116

Williams, D. *Probability with Martingales*. Cambridge Universtiy Press, 1991. 133

Yaari, M. E. Uncertain lifetime, life insurance, and the theory of the consumer. *The Review of Economic Studies*, 32(2):137–150, 1965. 56, 78

Zhang, J., Walter, G. G., Miao, Y., and Lee, W. N. W. Wavelet neural networks for function learning. *IEEE Transactions on Signal Processing*, 43(6):1485–1497, 1995. 21

Zălinescu, C. *Convex Analysis in General Vector Spaces*. World Scientific, 2002. 54

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.