THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Learning to Estimate: Bayesian Filtering with Deep Density Methods

Kasper Bågmark

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden 2026

Learning to Estimate: Bayesian Filtering with Deep Density Methods
Kasper Bågmark
Göteborg 2026
ISBN: 978-91-8103-410-3

Acknowledgements, dedications, and similar personal statements in this thesis, reflect the author's own views.

© Kasper Bågmark, 2026

Doktorshavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5867
ISSN 0346-718X
https://doi.org/10.63959/chalmers.dt/5867

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Cover: An unnormalized probability density of a periodic state process

Typeset with LaTeX
Printed by Chalmers Digitaltryck
Göteborg, Sweden 2026

till Dag och Ellen

Learning to Estimate: Bayesian Filtering with Deep Density Methods

Kasper Bågmark

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Bayesian filtering concerns the sequential estimation of the hidden state of a dynamical system from partial and noisy observations. Its central object is the conditional distribution of the hidden state given the available data, which provides both point estimates and a quantitative description of uncertainty. In nonlinear and non-Gaussian settings, this distribution is typically not available in closed form, and the construction of accurate and computationally feasible approximation methods becomes a central challenge, especially for high-dimensional systems.

This thesis studies Bayesian filtering with particular emphasis on density-based formulations when the underlying state is governed by a stochastic differential equation. We formulate the filtering problem through the evolution of conditional probability densities, described by stochastic and deterministic partial differential equations. Across the four appended papers, we develop methodologies for approximating these equations in high-dimensional settings. The proposed approaches draw on stochastic analysis, numerical analysis, and deep learning. They combine operator splitting, probabilistic backward representations, logarithmic transformations, and neural networks to approximate the conditional probability density. Theoretical convergence orders are established and verified numerically. The approaches are successfully demonstrated on nonlinear, high-dimensional, and partially observed stochastic differential equations.

Taken together, the papers in this thesis develop a framework for Bayesian filtering that combines probabilistic density representations with modern learning-based computational methods. The results indicate that these approaches can provide accurate and scalable alternatives to classical filtering methods for nonlinear and high-dimensional systems.

Keywords: Bayesian filtering, stochastic differential equations, partial differential equations, backward stochastic differential equations, density estimation, numerical methods, operator splitting, error estimates, learning-based approximations.

List of publications

This thesis is based on the work represented by the following papers:

I. **Bågmark, K.**, Andersson, A., Larsson, S. (2023). An energy-based deep splitting method for the nonlinear filtering problem. *Partial Differential Equations and Applications*, ***4**:14*, doi: 10.1007/s42985-023-00231-5.
II. **Bågmark, K.**, Andersson, A., Larsson, S., Rydin, F. (2024). A convergent scheme for the Bayesian filtering problem based on the Fokker–Planck equation and deep splitting. Submitted, preprint available at *arXiv:2409.14585*, doi: 10.48550/arXiv.2409.14585.
III. **Bågmark, K.**, Andersson, A., Larsson, S. (2025). Nonlinear filtering based on density approximation and deep BSDE prediction. Submitted, preprint available at *arXiv:2508.10630*, doi: 10.48550/arXiv.2508.10630.
IV. **Bågmark, K.**, Rydin, F. (2025). High-dimensional Bayesian filtering through deep density approximation. Submitted, preprint available at *arXiv:2511.07261*, doi: 10.48550/arXiv.2511.07261.

Additional papers not included in this thesis:

V. **Bågmark, K.**, Rydin, F. (2026). Neural likelihood surrogates for parameter inference via log-density PDE. *The 14th International Conference on Learning Representations Workshop AI&PDE*.

Author contributions

I. Derived the method in collaboration, performed the analysis and the numerical experiments. Wrote the paper with guidance and feedback from coauthors.

II. Performed the error analysis, and performed the numerical experiments together with coauthor. Wrote the paper with guidance and feedback from coauthors.

III. Performed the error analysis, and performed the numerical experiments. Wrote the paper with guidance and feedback from coauthors.

IV. Initiated the project, came up with the ideas, performed the derivations, wrote the paper independently with support from coauthor and advisor. Performed the experiments in collaboration with coauthor.

Acknowledgements

Acknowledging everyone who helped me bring this thesis to completion could easily fill two pages on its own, but acknowledging everyone who helped bring me this far in life, to the extent they deserve, would require far more pages than this thesis contains.

If I look for the academic beginning of that story, I find it here, at this department. Before becoming a PhD student here I had already spent both my bachelor's and master's years here. During my bachelor's thesis I was certainly not the ideal student (as my co-author and friend Emil can confirm) though never for lack of interest, only for lack of discipline. While writing that thesis I also took a course in partial differential equations with Mohammad. It was one of my favorite courses up to that point, and from then on I no longer had much doubt that a PhD was what I wanted to pursue. In the years that followed I took courses in stochastics, imagining for a while that I might become a quant or work in mathematical finance, but it was really the PDE courses, and especially the bridge between PDEs and stochastics, that kept pulling me back in. Taking my second PDE course with Axel made my growing interest in numerics impossible to ignore, and having such an inspiring teacher meant a great deal to me. In the end, I wrote my master's thesis with Annika on the simulation of random fields, a topic I found deeply fascinating, and through that work I received excellent mentorship in developing the skills needed to pursue a PhD. Annika guided me toward independence in the best possible way, and I think that has shaped much of how I have learned and carried out research during my PhD years as well.

In Annika, I have seen how determination, hard work, and strong leadership can truly result in an excellent research environment. Even though we have not written any papers together I have greatly valued the domain expertise and community knowledge she has brought, which has complemented the guidance from Adam and Stig.

To my supervisors and co-authors, Adam and Stig, I am grateful in more ways than I can easily express. What they both provide, to a degree that surpasses any supervision I have heard others describe, is time. At a moment's notice they are willing to look into whatever I need help with, answer questions, take meetings, switch gears, and prioritize my work above almost everything else. The time and care they have given so freely has meant an enormous amount to me. Alongside this, they both bring a level of perfectionism that I admittedly struggled with in my first papers, and sometimes still do. Both my patience and my confidence would often run out long before all the possible improvements to the story, the notation, or the exposition had been exhausted.

Yet in many ways this is the quality I think I will value most in future collaborators. It has made me strive for a standard that they would approve of, and that is not an easy standard to meet. I suspect I will compare many future collaborations to this one. They are also my friends: people I truly trust, people whose support I have never doubted, and people I very much want to make proud.

I want to thank Moritz for somehow ending up as the honorary extra wheel in my supervisory setup, and for carrying that role with great patience and very good humor. I have truly appreciated our research discussions, but just as much all the conversations about languages, strange jokes, and the kind of nonsense that makes academic life much more fun.

During my almost six years as a PhD student at this department, I have had the good fortune of receiving thoughtful advice and support from three different managers: Irina, Aila, and Serik. I am especially grateful to Irina for always being in my corner when important decisions has to be made. Beyond being my manager, she also joined Umberto and me for two years in the recruitment of new PhD students. It was often a stressful task, but also a rewarding one, and I am very glad to have shared it with two such competent and complementary colleagues.

Two doors down from my office sits Marija, the PhD students' best friend, and certainly one of the people here to whom I owe the most. She claims to enjoy helping with grant writing, which I still find hard to believe, but her skill in that area is such that no sensible person would ever turn down her help. She has mentored me in many ways, both in life and in my career, and I value her advice enormously. I consider her a friend with whom I want to share life's news (and I suspect I am very far from alone in feeling that way).

During my final year at the department I have had the feeling that I received help from almost the entire Operations Support Division. Ai-Linh in particular has had to deal with what, to me at least, felt like an endless stream of complicated calculations involving budgets, overhead, and other such mysteries. I am very grateful for all that help. I would also like to thank Julia, Mia, Frida, Jeanette, Helena, Jovan, Marie, Pernilla, Fia, Setta, and Ulf. I want to thank Anders, my examiner, for the trust and confidence you have shown me.

I am deeply grateful for the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation, whose importance to this community of researchers can hardly be overstated. Beyond funding my PhD, WASP has had a profound impact on my professional development. Through this network I have found collaborators and friends, received job opportunities, and had the chance to develop my leadership skills by leading a student-led cluster of PhD students. I would especially like to thank my now mostly graduated colleagues Anton, Amanda, Xixi, and Joel for helping create such a stimulating and enjoyable environment around probabilistic modeling.

In 2024, Moritz, Ruben, and I attended the Dynstoch workshop in Kiel and decided that Gothenburg should host the workshop. This turned out to be a slightly optimistic idea. Ruben, Moritz, and I set out enthusiastically and eventually convinced Annika, Umberto, and Pierre to join us. After seeing the modest and very pleasant edition in Kiel, we imagined that we could aim for something similar, perhaps even simpler. Instead, we discovered that hosting the same kind of workshop in Gothenburg would cost roughly three to four times as much, ironically just across the ferry route from Kiel. In the end, with support from Wilhelm och Martina Lundgrens Vetenskapsfond, Kungliga Vetenskapsakademien, and Wenner-Gren Stiftelserna, it became possible. Organizing this workshop has been both demanding and deeply rewarding, and I am grateful to all my co-organizers for making it happen. I am especially grateful to Ruben, who never gave up on the idea and kept letting me believe, one more time, that we would somehow manage to pull it off.

Apart from my supervisors Adam and Stig, I have also had the opportunity to work with Filip, Zheng, Moritz, Frank, and Daniele on projects and proposals. I am grateful to all of them for interesting discussions and enjoyable collaborations. A particularly warm thank you goes to Filip who began as a master's thesis student I supervised, but has since grown into a collaborator I value immensely, with expertise very complementary to my own. I do not say this lightly: working with Filip has been one of the most rewarding parts of my (still rather short) career.

During my time at the department I have primarily shared office with Johan and Oskar, and for shorter periods also with Erik and Henrik. Johan, thank you for always asking fundamental and interesting questions about everything from preschool pick-up logistics to distributions of distributions. Oskar, as one of my closest friends (both during our PhD years and long before that) I want to thank you for putting up with my rather extensive need for control and for generously allowing me to have opinions on nearly every aspect of your life. I am sorry about that (though apparently not sorry enough to stop). I am deeply grateful for all the laughter and all the inside jokes we have shared. Whether an acknowledgments section is really the right place for 'Bron' or 'Favoriten', I am still not entirely sure.

As the years have passed, I have found more and more friends among my colleagues at the department, and it truly saddens me that I am now the one finishing. I find it hard to imagine a more rewarding work environment, both in terms of the work itself and the people who make it what it is. I want to especially mention PhD student colleagues such as Erik, Per, Sebastian, Petar, Michael, and Gustav, who finished before me, as well as Selma, Anna[2], Mathis, Gijs, Niki, Henrik, Elias, Jenny, Isac, Mika, Linnea, Albert, Malin, Joseph, Philipp, Robin, Victor, Rickard, Lucia, Julia, Oskar, and Lotta, whom I managed to beat to the finish line. I would also like to thank Akash, Sagy, Helga, Axel[2], Ottmar, Jeff, Philip, Martin, and David for guidance and collegial friendship.

I want to thank Björn and Ioanna for being two of my closest friends at the department; you both mean a great deal to me. And Ruben, thank you for answering, or at least enduring, my endless stream of questions on everything from splitting operators and Microsoft Forms to cheap wine and boats. Few people move that naturally between

numerical analysis, administration, and questionable lifestyle advice. I am so grateful that I got to know you and I have a hard time imagining my final two years as a PhD student being nearly as fun or as meaningful otherwise.

I am also deeply grateful for all the kind conversations, cheerful greetings, and encouragement I have received from the many people not mentioned explicitly here.

Beyond the department, there are also people whose influence reaches much further back. I am fairly certain that I would not have ended up here without my elementary school teacher Linda. She taught mathematics and natural sciences and was very good at inspiring interest in them. At a time when I had more or less decided to pursue technical studies without much emphasis on science, she strongly encouraged me to consider Bäckängsgymnasiet and the natural sciences track. Thankfully, I did, and that choice set me on the path that eventually led to this thesis.

Just as importantly, that choice led me to an amazing class and to some of the closest friends I still have today. Ivana, Emil, and Georg have all already defended their PhDs. John and Emelie are, after my family, the most important people in my life. Time and again throughout my adult life, they have helped pick me up, and I am immensely grateful to have them.

I want to thank my parents and my brother for supporting me in every endeavor. In countless ways, you have helped me as I have tried to figure out life through all its ups and downs.

Ellen, you give me so much love, support, and trust, and every day I get to spend with you feels like winning the lottery. Thank you for being my best teammate and for always being by my side. I will always be by yours.

Dag, my child. I wish you all the best in this world, and I will do my very best to give it to you. I am far from perfect, but I want you to know that I love you more than anything in this world.

Contents

1 Introduction

Imagine a vessel moving through a coastal region during a storm. Its intended route is not completely fixed, and over time its motion is disturbed by wind, waves, and other unpredictable effects. The surrounding waters are not entirely safe. Shallow areas and dangerous rocks lie hidden along parts of the coastline, so even a modest deviation from course may place the vessel at risk before it reaches the harbor. From the shore, however, the situation is only partially visible. A number of coastal radar stations pick up noisy signals. Each gives an imperfect indication of where the vessel may be. Some readings suggest that it is close to a particular part of the coast, while others are less conclusive. These may be distorted by distance, weather, or the angle from which the vessel is observed. In Figure 1.1 we see a simplified illustration of this scenario.

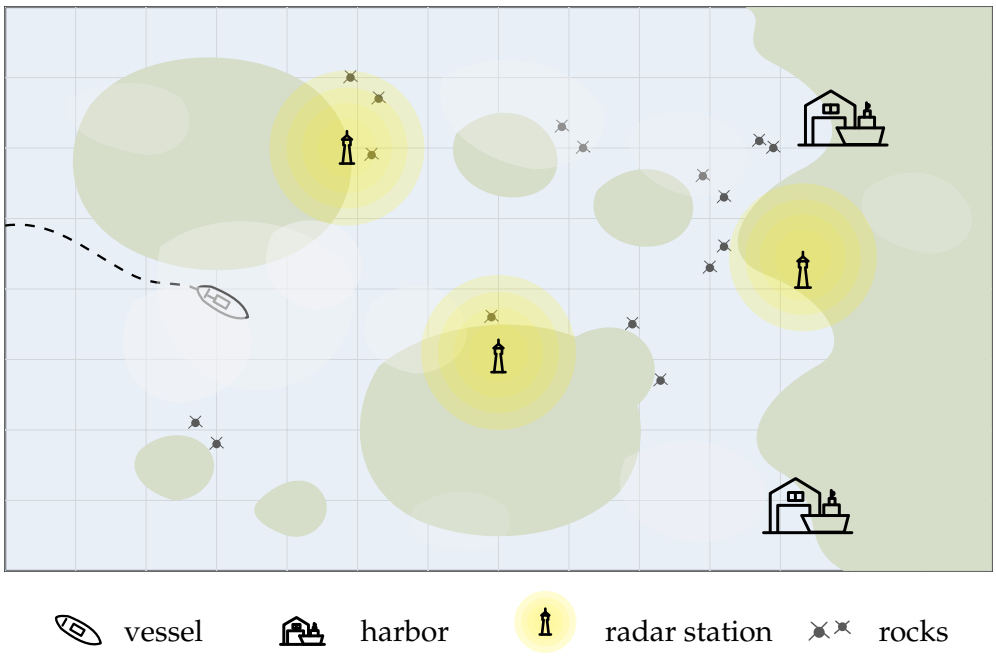


Figure 1.1: A vessel moves through uncertain coastal waters, while coastal radar stations provide partial and imperfect information about its location.

Suppose now that we are working in a sea rescue center, following the situation as these measurements arrive over time. The solution is to combine the incoming observations with a physical description of the vessel's movement and a statistical description of the sensing capability. In this way, we seek to infer where it is likely to be and whether it may be approaching danger. Since uncertainty is unavoidable in the unpredictable sea and with inexact measurements, the natural goal is not a single estimated trajectory. Rather, it is a probability distribution describing where the vessel may be at a given time, conditioned on the information collected so far. This evolving distribution is the central object in Bayesian filtering.

Although this example is easy to picture, it already contains the essential ingredients of Bayesian filtering. There is some underlying state S of the system that we would like to know, but cannot observe directly. Instead, we receive observations O, which provide only partial and noisy information about that state. The role of Bayesian filtering is to combine these observations with prior knowledge and with a model for how the system evolves. The aim is to update our uncertainty over time. As new information arrives, our belief about the current state is revised. This leads to a sequential probabilistic description of what may be happening in the system.

Problems of this kind arise in many different settings. In navigation and surveillance, one may wish to track the position of a vessel, aircraft, or satellite from incomplete sensor data (Bar-Shalom et al., 2001; Blackman and Popoli, 1999). In robotics, filtering is used for localization and mapping when a robot must infer its position from noisy measurements of its surroundings (Thrun et al., 2005). In finance and signal processing, hidden variables are often estimated sequentially from indirect observations (Brigo and Hanzon, 1998; Date and Ponomareva, 2011). In meteorology and oceanography, observational data are combined with large dynamical models in order to improve forecasts (Evensen, 1994, 2003; Apte et al., 2008). Related ideas also appear in biology and medicine, where one may seek to infer unobserved physiological or population-level processes from partial measurements (Ricciardi and Sacerdote, 1979; Kamino et al., 2023). While the applications differ greatly in scale and interpretation, they share the same underlying challenge. One seeks to estimate an evolving and only partially observed system in the presence of uncertainty.

This thesis studies Bayesian filtering. The specific emphasis of the works presented here is on computational methods and scalability for the filtering problem. In addition, the thesis focuses on probabilistic descriptions of uncertainty rather than only pointwise estimates.

A brief historical overview

The study of filtering has a long history, with roots in control theory, signal processing, statistics, and probability. A decisive early breakthrough came with the work of Wiener on linear estimation and prediction (Wiener, 1949). Closely related and independent contributions were also made by Kolmogorov on interpolation and extrapolation of stationary random sequences (Kolmogorov, 1941). Later, the state-space perspective introduced by Kalman (Kalman, 1960) marked a major development in the field, although related recursive estimation ideas had appeared slightly earlier in work by Swerling (Swerling, 1959). In his seminal 1960 paper, Kalman showed that for linear dynamical systems with Gaussian noise, the evolving uncertainty about the hidden state can be propagated recursively through a finite-dimensional set of equations (Kalman, 1960). Shortly thereafter, Kalman and Bucy developed the corresponding continuous-time theory (Kalman and Bucy, 1961). These results were mathematically elegant, computationally tractable, and highly influential in applications. They established filtering as a central topic in modern applied mathematics.

Beyond the linear-Gaussian setting, however, the filtering problem becomes substantially more difficult. In nonlinear and non-Gaussian models, the conditional distribution of the hidden state typically no longer admits a closed finite-dimensional description. In the case of continuous-time observations, one is instead led to nonlinear filtering equations such as the Kushner–Stratonovich equation, the Zakai equation, and the Fujisaki–Kallianpur–Kunita equation. These describe the evolution of the conditional distribution, or related unnormalized quantities, in infinite-dimensional form (Stratonovich, 1960; Wonham, 1964; Kushner, 1964; Zakai, 1969; Fujisaki et al., 1972). Closely related to this development is the Kallianpur–Striebel formula (Kallianpur and Striebel, 1968). It provides a Bayes-type representation of the conditional distribution and has become one of the foundational tools of nonlinear filtering theory. This line of development drew on contributions from several researchers and provided the rigorous mathematical foundation for continuous-time nonlinear filtering. At the same time, it made clear that exact filtering is rarely available outside a small class of special models. Effective approximation methods are therefore essential in practice.

A large part of the subsequent literature has therefore focused on approximation and computation. One line of work extends Gaussian filtering ideas beyond the classical Kalman setting. This leads, for example, to the extended Kalman filter and, later, to ensemble Kalman methods. In particular, the ensemble Kalman filter introduced by Evensen (1994) and further developed by Houtekamer and Mitchell (1998) and Burgers et al. (1998) became highly influential in data assimilation. This was especially true in meteorology and

oceanography, where high-dimensional dynamical systems make explicit covariance propagation infeasible. Another major development was the emergence of particle filters, which represent the filtering distribution by weighted samples rather than by Gaussian summaries. The bootstrap filter of Gordon et al. (1993) is often regarded as a landmark contribution. Particle filtering was subsequently developed in several important directions, including the Monte Carlo filter of Kitagawa (1996) and the auxiliary particle filter of Pitt and Shephard (1999). These methods greatly expanded the range of nonlinear and non-Gaussian problems that could be treated computationally.

More recently, research has continued along several directions, including improved particle methodologies, high-dimensional filtering strategies, and learning-based approaches for sequential inference (Naesseth et al., 2019; Finke and Thiery, 2023; Chopin et al., 2023; Luk et al., 2024; Bao et al., 2024). At the same time, the fundamental computational difficulties of nonlinear filtering remain, especially in high-dimensional settings where both classical Gaussian approximations and particle-based methods may become inadequate or prohibitively expensive. Particle filters nevertheless remain a central reference methodology for nonlinear and non-Gaussian filtering, both because of their generality and because they provide a natural benchmark against which new approaches can be assessed. The present thesis is situated within this broader development, with particular emphasis on Bayesian filtering through density-based formulations and modern computational methodology.

1.1 Problem formulation

In the setting considered in this thesis, there is an underlying state process S evolving in continuous time according to a stochastic dynamical model. Observations O are collected either continuously or at discrete time points. The basic task of Bayesian filtering is to estimate the current state of the system from the information provided by these observations, illustrated in Figure 1.2. At first sight, one might think of this as a problem of producing a best estimate or a most likely trajectory. However, in many situations this is not sufficient. The available data are partial and noisy. The system itself evolves randomly, and several different states may be compatible with the same observation history. What is needed is therefore not only an estimate of where the system is, but also a quantitative description of the remaining uncertainty.

For this reason, the central mathematical object in filtering is the conditional distribution of the hidden state S at the current time t_k given the observations O collected at time points $t_1, \ldots, t_k$. This marginal distribution represents the

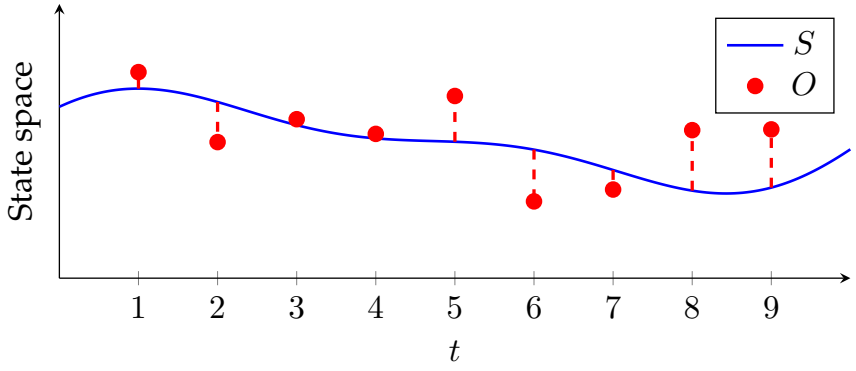


Figure 1.2: A simple illustration of the filtering setting considered in this thesis. A hidden state process S evolves continuously in time, while noisy observations O are collected at discrete time points and are used to infer the state of the system.

probabilistic information that can be inferred about the system from the model and the data. It allows one not only to extract point estimates such as conditional means, or modes of the distribution, but also to quantify uncertainty. It also allows one to assess the likelihood of rare but important events and to compute other quantities of interest relevant for prediction and decision-making. In this sense, the conditional distribution contains a far richer description than any single representative trajectory could.

- Goal: Infer $\mathbb{P}(S_{t_k} \mid O_{1:k})$, the distribution of S at time t_k given k observations $O_{1:k} := (O_1, \ldots, O_k)$.

A particularly important viewpoint, and one that is central throughout this thesis, is to represent this conditional distribution through its probability density whenever such a density exists. In this formulation, the filtering problem becomes the task of computing or approximating an evolving density over the state space. This density is updated sequentially as new observations arrive. Between observations, it evolves according to a partial differential equation of Fokker–Planck type. At observation times, it is updated through Bayes' rule. This density-based perspective provides a natural and expressive probabilistic description of uncertainty. It also gives direct access to the full shape of the filtering distribution. At the same time, it leads to substantial analytical and computational challenges.

These challenges are particularly severe in nonlinear and high-dimensional settings. Except in special cases, the filtering distribution is not available in closed form, and one must instead rely on numerical approximation. Classical methods such as the extended Kalman filter often depend on local Gaussian approximations. These may become inaccurate in strongly nonlinear regimes

or when the underlying distribution is far from Gaussian. Particle filters are conceptually flexible and applicable to a broad range of nonlinear and non-Gaussian models, but may require very large numbers of particles in high dimensions. In such settings, the information carried by the particles tends to deteriorate rapidly. After reweighting, only a small number of particles may contribute meaningfully, while the rest carry negligible weight. To maintain a satisfactory representation of the distribution, the number of particles may then need to grow extremely quickly with dimension. This leads to a severe computational burden. This phenomenon is one manifestation of the curse of dimensionality. Ensemble Kalman methods offer a scalable alternative in some settings, but their Gaussian structure can limit their ability to represent more complicated distributions. Taken together, this creates a need for new filtering methodologies. Such methods should capture genuinely nonlinear and non-Gaussian behavior while remaining computationally feasible in high-dimensional state spaces.

The present thesis is concerned with the development and analysis of such methods. The emphasis is on density-based formulations of Bayesian filtering, and on mathematical, computational, and deep learning-based approaches that make these formulations useful beyond low-dimensional benchmark problems.

1.2 Aims

The overall aim of this thesis is to develop and analyse methods for Bayesian filtering in nonlinear and high-dimensional settings, with particular emphasis on density-based formulations. More specifically, the thesis pursues the following objectives:

- formulate Bayesian filtering problems in terms of conditional distributions and densities in a way that retains their probabilistic properties,
- develop numerical and learning-based methods for approximating these distributions in settings where exact solutions are unavailable,
- design such methods to capture nonlinear and non-Gaussian structure while remaining computationally feasible in high-dimensional settings,
- investigate theoretical properties of the proposed methods, including approximation and convergence aspects where possible.

Outline of the thesis

This thesis is organized as follows. In Chapter 2, we introduce the class of state space models considered throughout the thesis. We also review the basic filtering concepts that underlie all four papers. This includes the general prediction-update structure, the linear Gaussian case leading to the Kalman filter, and a brief overview of classical approximation methods for nonlinear filtering. These serve as important points of comparison in the appended works.

Chapter 3 provides the mathematical background needed for the density-based methodologies developed later in the thesis. In particular, it introduces the stochastic differential equation framework for the hidden state process. It also presents the associated Kolmogorov and Fokker–Planck equations, together with the Feynman–Kac representations that connect the probabilistic and partial differential equation viewpoints. These tools form the basis for both the analytical developments and the computational methods used in Papers I–IV.

In Chapter 4, we present the main methodological ideas that unify the appended papers. The chapter introduces the density-based problem formulation for the filtering problem, together with the operator splitting and optimization-based perspectives used to construct the proposed approximation methods. In this way, it serves as a bridge between the general background material and the specific contributions of the papers.

Finally, Chapter 5 provides a summary of the four appended papers and explains how they relate to one another. Paper I introduces a learning-based splitting method for the continuous-observation setting through the Zakai equation. Paper II develops the corresponding discrete-observation framework based on the Fokker–Planck equation and includes a convergence analysis of the proposed scheme. Paper III studies the same discrete filtering problem through a probabilistic representation based on forward backward stochastic differential equations and establishes theoretical error results for the resulting approximation. Paper IV extends and compares the density-based methods from Papers II–III in high-dimensional settings, with particular emphasis on logarithmic density formulations and computational performance.

2 State space models and filtering

This chapter introduces Bayesian filtering for the class of models considered in the thesis, and focuses in particular on the mathematical tools on which the methods in Papers I–IV rely. We start by introducing general state spaces in Section 2.1, where the states are hidden from us and observations are provided sequentially in time. In Section 2.2 we continue by describing the Kalman filter which is the analytical solution to the filtering problem in the linear case. Finally, in Section 2.3 we include some classic filtering algorithms used for the nonlinear filtering scenario. We begin by introducing some notation used throughout the thesis.

Notation

Let d and d' be positive integers. Throughout the thesis, $\mathbb{R}^d$ denotes the d-dimensional Euclidean space and $\mathbb{N}$ the set of positive integers. We use S to denote the hidden state variable, which is $\mathbb{R}^d$-valued, and O to denote the observation process, which is $\mathbb{R}^{d'}$-valued. The observation process can either be continuous in time, as in Paper I, or discrete in time, as in Papers II–IV. We consider a time domain $[t_0, T]$, where we let the initial time be fixed as $t_0 = 0$, and $T >$ denotes the final time $T > 0$. Furthermore, we let $K \in \mathbb{N}$ be the number of observations, where $t_1, \ldots, t_K \in [0, T]$ are the observation times in increasing order.

For a discrete observation sequence, we write $O_{1:k} = (O_1, \ldots, O_k)$, $k = 0, \ldots, K$, with the convention $O_{1:0} = \emptyset$. When convenient, we identify such a sequence with an element of $\mathbb{R}^{d' \times k}$. We commonly use capital letters for random variables and lowercase letters for deterministic variables or realizations.

Thus, if Ω is the underlying sample space and $\omega \in \Omega$, then a realization of the observation sequence is written $o_{1:k} = O_{1:k}(\omega)$.

The multivariate Gaussian distribution is denoted by $\mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. When we write $\mathcal{N}(x \mid \mu, \Sigma)$, this refers to the corresponding Gaussian density evaluated at $x \in \mathbb{R}^d$. We also reserve Φ for the standard Gaussian cumulative distribution function.

We let $C^n(\mathbb{R}^d; \mathbb{R})$ denote the space of n times continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$, and $C^{1,2}([0,T] \times \mathbb{R}^d; \mathbb{R})$ the space of functions that are once continuously differentiable in time and twice continuously differentiable in space. For a differentiable function φ, we write $\nabla\varphi$ for its spatial gradient in the scalar-valued case, and $D\varphi$ for its Jacobian matrix in the vector-valued case. We use $\|\cdot\|$ to denote the Euclidean norm, and I_d the identity, in $\mathbb{R}^d$.

We use $\mathbb{E}[\cdot]$ for the expectation and $\mathbb{E}[\cdot \mid \cdot]$ for the conditional expectation. Throughout this thesis, $p(\cdot \mid \cdot)$ denotes a conditional probability density function with respect to the Lebesgue measure on $\mathbb{R}^d$ (or $\mathbb{R}^{d'}$ for observations), whenever such a density exists. Expressions such as $p(S_{t_k} = x \mid \cdot)$ are understood as shorthand for the conditional density of S_{t_k} evaluated at x.

The signal process S is, from Section 3.1 onwards, driven by a Brownian motion B. In later probabilistic representations we also use an auxiliary process X, driven by a Brownian motion W, to keep this notation separate from that of the signal. The process X is also used when no hidden signal is considered, to emphasize its different role from S.

2.1 State space models

A convenient way to represent the filtering problem is through the Markovian state space model shown in Figure 2.1. The model consists of a hidden state process S, evolving forward in time, and an observation process O, which provides sequential partial information about the state. Both components are described through conditional distributions: one for the state dynamics and one for the observation mechanism.

A key feature of the model is the Markov property of the state process. In the graphical representation, this is reflected by the absence of arrows from S_{t_j} to S_{t_k} for $j < k-1$. More precisely, conditional on $S_{t_{k-1}}$, the state S_{t_k} is independent of $(S_0, \ldots, S_{t_{k-2}})$. This structure underlies the recursive form of the filtering equations. The precise mathematical notion of S used throughout the thesis is introduced later in Section 3.1.

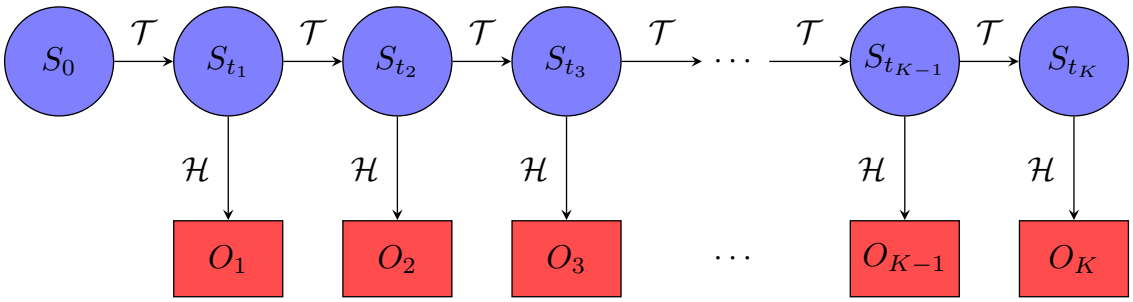


Figure 2.1: The state space with continuous state S propagating forward with random dynamics denoted by $\mathcal{T}$, and discrete observations O measured through a measurement function $\mathcal{H}$.

In sequential probabilistic estimation problems, there are three central conditional distributions one typically aims to estimate: the prediction distribution, the filtering distribution, and the smoothing distribution.

1. **Prediction:** Estimate the prediction density $p(S_{t_k} \mid O_{1:j})$ with $k \in \{1, \ldots, K\}$ and $j \in \{0, \ldots, k-1\}$. In other words, estimate a future state given observations from the past.

2. **Filtering:** Estimate the filtering density $p(S_{t_k} \mid O_{1:k})$ for $k \in \{1, \ldots, K\}$, i.e., estimate the current state given observations up to the current time.

3. **Smoothing:** Estimate the smoothing density $p(S_{t_k} \mid O_{1:K})$ for $k \in \{1, \ldots, K\}$, i.e., estimate a past state given observations up to the final time.

To illustrate the distinction, we return to the vessel example from Chapter 1. Suppose that at time t_k we have received radar observations up to time t_k, and that another observation will be recorded at a later time t_{k+1}. Then the *prediction* problem consists of estimating where the vessel will be at time t_{k+1} given the currently available observations up to time t_k. The *filtering* problem consists instead of estimating where the vessel is at the present time t_k using the observations collected up to that same time. Finally, the *smoothing* problem asks where the vessel was at some earlier time $t_j < t_k$, now using the additional information contained in observations up to the later time t_k. In this way, prediction concerns the future, filtering the present, and smoothing the past.

Generally, in sequential estimation theory, one solves these problems in the order given above. Often, to perform smoothing at a time point t_k with $k < K$, one first carries out prediction and filtering up to the final time t_K, and then

applies backward techniques to obtain the smoothing estimate. In this thesis, we do not discuss smoothing in any great detail, but instead focus on the filtering problem.

The recursive structure is essential in applications where observations arrive sequentially and decisions must be updated in real time. Returning again to the vessel example, a rescue center cannot wait until all future radar measurements have been collected before assessing whether the vessel is approaching shallow water. Instead, each new observation must immediately be incorporated into the current uncertainty description. The filtering equations provide exactly such a mechanism: the prediction step propagates the current uncertainty forward according to the motion model, while the update step corrects this prediction using the newly arrived observation. This recursive prediction-update cycle is one of the central structural features of Bayesian filtering.

Similarly to smoothing, to conduct filtering, we need to solve the prediction problem. For this reason we introduce the filtering equations, consisting of a prediction step and an update step. We assume a prior distribution on S, at time $t_0 = 0$, such that $S_0 \sim p_0$ for some fixed distribution p_0. The equations are, initialized by $p(S_0) = p_0$ and recursively for $k = 0, \dots, K-1$, given by

$$p\big(S_{t_{k+1}} = x \,\big|\, O_{1:k}\big) = \int_{\mathbb{R}^d} p\big(S_{t_{k+1}} = x \,\big|\, S_{t_k} = y\big)\; p(S_{t_k} = y \,|\, O_{1:k})\,\mathrm{d}y, \tag{2.1}$$

$$p(S_{t_k} = x \,|\, O_{1:k}) = \frac{p(O_k \,|\, S_{t_k} = x)\; p(S_{t_k} = x \,|\, O_{1:k-1})}{\int_{\mathbb{R}^d} p(O_k \,|\, S_{t_k} = y)\; p(S_{t_k} = y \,|\, O_{1:k-1})\,\mathrm{d}y}. \tag{2.2}$$

We note that the prediction "equation" (2.1) and update "equation" (2.2) are, strictly speaking, formulas; however, they are commonly referred to as equations in the literature, and we retain this terminology. Looking at the prediction density at time t_{k+1}, the left hand side of (2.1), we see that it depends on the filtering density at the previous time step t_k, together with the transition density from S_{t_k} to $S_{t_{k+1}}$ denoted by $p(S_{t_{k+1}} \mid S_{t_k})$. The update step is Bayes' rule, where the denominator simplifies to $p(O_k \mid O_{1:k-1})$.

2.2 Linear filtering

We now consider the special case in which the state dynamics and observation model are linear, and all uncertainty enters through Gaussian noise. In this setting, the filtering problem admits a closed-form recursive solution, given by the Kalman filter.

The key reason is that Gaussian distributions are preserved by the two steps in the filtering recursion. First, if the current filtering distribution is Gaussian and the state evolves through a linear transformation with additive Gaussian noise, then the prediction distribution remains Gaussian. Second, if the observation model is linear with additive Gaussian noise, then Bayes' update again yields a Gaussian distribution. Hence, starting from a Gaussian initial distribution, all subsequent prediction and filtering distributions remain Gaussian and are therefore completely characterized by their means and covariance matrices.

This recursive Gaussian structure is made precise by the Kalman filter (Kalman, 1960), named after Rudolf E. Kálmán. The corresponding continuous time version, with linear state and observation dynamics, was derived in Kalman and Bucy (1961). We refer to Särkkä and Svensson (2023) for a comprehensive treatment of the Kalman filter and its extensions.

Method 2.2.1: Kalman filter

Assume a linear Gaussian state space model of the form

$$\begin{aligned} S_{t_{k+1}} &= A_k S_{t_k} + \xi_k, & \xi_k &\sim \mathcal{N}(0, C), \\ O_k &= H_k S_{t_k} + \eta_k, & \eta_k &\sim \mathcal{N}(0, D), \end{aligned}$$

with initial distribution $S_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$. The filtering distribution remains Gaussian, with mean μ_k and covariance matrix Σ_k, at all times,

$$p(S_{t_k} \mid O_{1:k}) = \mathcal{N}(\mu_k, \Sigma_k).$$

The parameters are updated recursively, together with prediction parameters $(\mu_{k|k-1}, \Sigma_{k|k-1})$ and auxiliary parameters (ν_k, K_k) as follows.

Prediction step.

$$\begin{aligned} \mu_{k|k-1} &= A_k \mu_{k-1}, \\ \Sigma_{k|k-1} &= A_k \Sigma_{k-1} A_k^\top + C. \end{aligned}$$

Update step.

$$\begin{aligned} \nu_k &= O_k - H_k \mu_{k|k-1}, \\ K_k &= \Sigma_{k|k-1} H_k^\top (H_k \Sigma_{k|k-1} H_k^\top + D)^{-1}, \\ \mu_k &= \mu_{k|k-1} + K_k \nu_k, \\ \Sigma_k &= (I - K_k H_k) \Sigma_{k|k-1}. \end{aligned}$$

2.3 Nonlinear filtering

Now, if instead the forward dynamics of S, denoted $\mathcal{T}$ in Figure 2.1, are nonlinear in the state variable, we cannot obtain a closed-form solution except in special cases (see the Beneš filter (Beneš, 1981)). Instead, we must apply approximate methods to obtain good estimates of the filtering distribution. This is a classical problem that has given rise to the field of nonlinear filtering, with a rich theory (Bain and Crisan, 2009; Särkkä and Svensson, 2023).

In Chapter 4, we discuss a problem formulation that represents the *partial differential equation approach* to the filtering problem. However, more *classical approaches* to estimating the filtering distribution include Kalman-based approximations, particle-based methods, or combinations of the two (Särkkä and Svensson, 2023). Next, we describe three important classical methods that we use as benchmarks in Papers I–IV.

The simplest approach, yet highly effective in many scenarios, is the extended Kalman filter, where the nonlinearity is approximated by a first-order Taylor expansion, yielding an approximate Gaussian distribution that evolves according to Kalman-type recursions. This filter is computationally efficient since it only requires multiplication and inversion of $d \times d$ matrices (which is comparatively cheap to do in dimensions below 10^3), as well as evaluations of the derivatives of the forward and measurement dynamics (e.g. $\mathcal{T}$ and $\mathcal{H}$ in Figure 2.1).

Method 2.3.1: Extended Kalman filter

Assume a nonlinear state space model of the form

$$\begin{aligned} S_{t_{k+1}} &= b(S_{t_k}) + \xi_k, & \xi_k &\sim \mathcal{N}(0, C), \\ O_k &= h(S_{t_k}) + \eta_k, & \eta_k &\sim \mathcal{N}(0, D), \end{aligned}$$

with initial distribution $S_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$. The extended Kalman filter approximates the filtering distribution by a Gaussian,

$$p(S_{t_k} \mid O_{1:k}) \approx \mathcal{N}(\mu_k, \Sigma_k),$$

whose parameters are updated by locally linearizing the model coefficients b and h evaluated in the previous mean. Similarly to the ordinary Kalman filter it is updated as follows.

Prediction step. Let $A_k = Db(\mu_{k-1})$ denote the Jacobian of b evaluated at μ_{k-1}. Then

$$\begin{aligned}\mu_{k|k-1} &= b(\mu_{k-1}),\\ \Sigma_{k|k-1} &= A_k \Sigma_{k-1} A_k^\top + C.\end{aligned}$$

Update step. Let $H_k = Dh(\mu_{k|k-1})$ denote the Jacobian of h evaluated at the predicted mean. The updates are given by

$$\begin{aligned}\nu_k &= O_k - h(\mu_{k|k-1}),\\ K_k &= \Sigma_{k|k-1} H_k^\top (H_k \Sigma_{k|k-1} H_k^\top + D)^{-1},\\ \mu_k &= \mu_{k|k-1} + K_k \nu_k,\\ \Sigma_k &= (I - K_k H_k)\Sigma_{k|k-1}.\end{aligned}$$

The extended Kalman filter is computationally efficient and widely used in practice, but its performance may deteriorate in strongly nonlinear systems due to the local linearization.

Example 2.3.1: Extended Kalman filter in a bistable model

Consider the one-dimensional nonlinear state space model

$$\begin{aligned}S_{t_{k+1}} &= b(S_{t_k}) + \xi_k, \qquad \xi_k \sim \mathcal{N}(0, C),\\ O_k &= h(S_{t_k}) + \eta_k, \qquad \eta_k \sim \mathcal{N}(0, D),\end{aligned}$$

with observation times given by $t_k = k\tau$, $k = 0, \ldots, K$, and

$$b(x) = x + \tau(ax - cx^3).$$

We apply the extended Kalman filter, approximating the filtering distributions by Gaussians $\mathcal{N}(\mu_k, \Sigma_k)$. We also assume $S_0 \sim \mathcal{N}(0.1, 1)$, and initialize the extended Kalman filter accordingly with $\mu_0 = 0.1$ and $\Sigma_0 = 1$ and follow the prediction and update steps as in Method 2.3.1.

In Figure 2.2, we choose $a = 9$, $c = 1$, $\tau = 0.02$, $K = 50$, $C = 1$, and $D = 0.5$. The drift term $ax - cx^3$ induces a bistable behavior in the state dynamics, with two stable regions near $\pm\sqrt{a/c} = \pm 3$. In the figure, we consider the two observation functions $h(x) = x$ and $h(x) = |x|$.

In the first case, the observations identify the sign of the state, and the filter tracks the trajectory accurately. In the second case, the sign information is lost, and the Gaussian approximation may instead lock onto the wrong well. This illustrates a basic limitation of the extended Kalman filter in nonlinear and non-Gaussian settings.

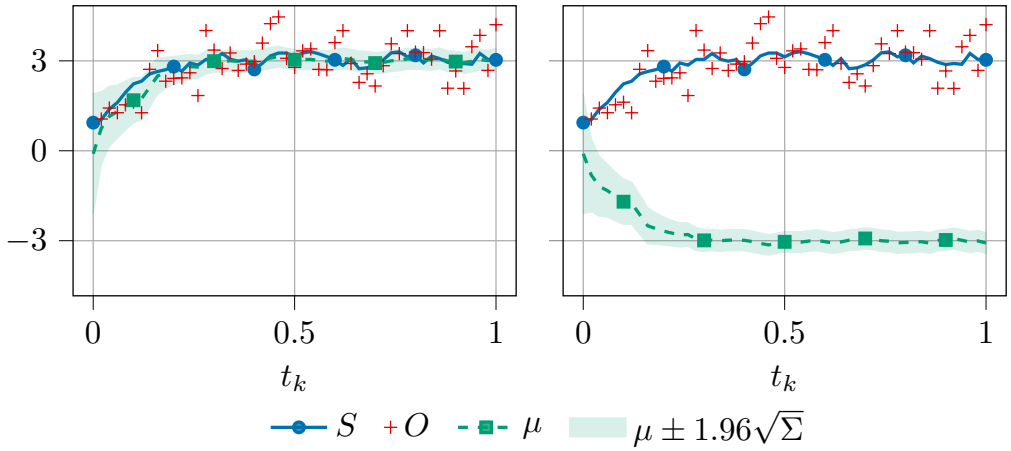


Figure 2.2: Illustration of the extended Kalman filter for a bistable state model with two different observation functions, $h(x) = x$ on the left, and $h(x) = |x|$ on the right. We illustrate the state, the observations, and the extended Kalman filter mean with a corresponding 95% confidence region.

Another approach is the ensemble Kalman filter. It is fundamentally different from the extended Kalman filter in that it represents the filtering distribution by an empirical ensemble, but it still shares important Gaussian assumptions. We refer to Evensen (1994) for the early work and to Burgers et al. (1998); Katzfuss et al. (2016) for work on the so called stochastic ensemble Kalman filter which we use in Paper IV. This method shares many similarities with sequential Monte Carlo methods, also known as particle filters. In all of the included papers we use a bootstrap particle filter as a benchmark (Gordon et al., 1993). Next we briefly describe both of these methods aimed at approximating the filtering density in the nonlinear setting.

Instead of modelling the prediction step directly through a deterministic approximation, we represent the distribution by particles (or ensemble members), representing the distribution, aimed at approximating the conditional distribution and behavior of the true signal S evolving over time. More precisely, we let a pair (X, W) denote a particle X and its weight W in the approximate distribution. To assimilate the conditional distribution of the signal S given observations O, we simulate M pairs $(X^{(i)}, W^{(i)})_{i=1}^{M}$ and define our approxi-

mation π by the empirical distribution

$$\pi = \sum_{i=1}^{M} W^{(i)} \delta_{X^{(i)}}, \tag{2.3}$$

where the weights sum to one, i.e., $\sum_{i=1}^{M} W^{(i)} = 1$. The prediction step of both methods, from step $k-1$ to step k, consists of collecting one sample from the conditional transition probability $p(X_k \mid X_{k-1}^{(i)})$, for every $X_{k-1}^{(i)}$, $i = 1, \ldots, M$. This transition density is designed to approximate, or in some cases equal, $p(S_{t_k} \mid S_{t_{k-1}})$. The methods differ mainly in how they handle the update step. The *bootstrap particle filter* employs a likelihood calculation by evaluating the likelihood of the observation given the obtained particles. In the *ensemble Kalman filter*, the update step instead consists of incorporating the information from the observation by evaluating the Kalman gain (present in the ordinary Kalman filter as K_k) obtained by the measurement function. See the respective method below for the complete methodology.

Method 2.3.2: Ensemble Kalman filter

Assume a nonlinear state space model of the form

$$\begin{aligned} S_{t_{k+1}} &= b(S_{t_k}) + \xi_k, & \xi_k &\sim \mathcal{N}(0, C), \\ O_k &= h(S_{t_k}) + \eta_k, & \eta_k &\sim \mathcal{N}(0, D), \end{aligned}$$

with initial distribution $S_0 \sim p_0$. The ensemble Kalman filter represents the filtering distribution at time t_k by an ensemble of particles with fixed uniform weights,

$$\pi_k(x) = \frac{1}{M} \sum_{i=1}^{M} \delta_{X_k^{(i)}}(x).$$

Prediction step. Given an ensemble $\{X_{k-1}^{(i)}\}_{i=1}^{M}$ approximating $p(S_{t_{k-1}} \mid O_{1:k-1})$, each particle is propagated according to the state dynamics,

$$X_{k|k-1}^{(i)} = b(X_{k-1}^{(i)}) + \xi_k^{(i)}, \qquad \xi_k^{(i)} \sim \mathcal{N}(0, C), \quad i = 1, \ldots, M.$$

Update step. To assimilate the observation O_k, the stochastic ensemble Kalman filter introduces perturbed observations

$$Y_k^{(i)} = h(X_{k|k-1}^{(i)}) + \eta_k^{(i)}, \qquad \eta_k^{(i)} \sim \mathcal{N}(0, D), \qquad i = 1, \ldots, M.$$

Based on the forecast ensemble, empirical covariances P_k^{xy} and P_k^{yy} between the particles and the predicted observations are computed, yielding an approximate Kalman gain

$$K_k = P_k^{xy} \, (P_k^{yy})^{-1}.$$

Each particle is then updated accordingly,

$$X_k^{(i)} = X_{k|k-1}^{(i)} + K_k\big(O_k - Y_k^{(i)}\big), \qquad i = 1, \ldots, M.$$

In the linear Gaussian setting, the ensemble Kalman filter converges to the classical Kalman filter as the ensemble size $M \to \infty$ (Evensen, 2003). In nonlinear problems, it provides an efficient Gaussian approximation of the filtering distribution, but perform poorly in strongly nonlinear regimes due to its reliance on linear updates and empirical covariances (Katzfuss et al., 2016).

Method 2.3.3: Bootstrap particle filter

Assume a nonlinear state space model, with $S_0 \sim p_0$, of the form

$$\begin{aligned} S_{t_{k+1}} &= b(S_{t_k}) + \xi_k, & \xi_k &\sim \mathcal{N}(0, C), \\ O_k &= h(S_{t_k}) + \eta_k, & \eta_k &\sim \mathcal{N}(0, D). \end{aligned}$$

The bootstrap particle filter represents the filtering distribution, initialized with uniform weights, by a weighted empirical measure

$$\pi_k(x) = \sum_{i=1}^{M} W_k^{(i)} \, \delta_{X_k^{(i)}}(x), \qquad \sum_{i=1}^{M} W_k^{(i)} = 1.$$

Prediction step. Given particles $\{(X_{k-1}^{(i)}, W_{k-1}^{(i)})\}_{i=1}^{M}$, each particle is propagated according to the state dynamics,

$$X_{k|k-1}^{(i)} = b(X_{k-1}^{(i)}) + \xi_k^{(i)}, \qquad \xi_k^{(i)} \sim \mathcal{N}(0, C), \quad i = 1, \ldots, M.$$

Update step. The particle weights are updated using the likelihood of the observation,

$$\widetilde{W}_k^{(i)} = W_{k-1}^{(i)}\, p(O_k \mid X_{k|k-1}^{(i)}),$$
$$W_k^{(i)} = \frac{\widetilde{W}_k^{(i)}}{\sum_{j=1}^{M} \widetilde{W}_k^{(j)}}, \qquad i = 1, \ldots, M.$$

Resampling. A new set of particles $\{X_k^{(i)}\}_{i=1}^M$ is sampled with replacement from $\{X_{k|k-1}^{(i)}\}_{i=1}^M$ according to the weights $\{W_k^{(i)}\}_{i=1}^M$, and the weights are reset to $W_k^{(i)} = \frac{1}{M}$.

We state the weight update explicitly because different particle filtering methods mainly differ in their choice of proposal distribution, weight update, and resampling strategy. The bootstrap particle filter provides a consistent Monte Carlo approximation of the filtering distribution, but typically suffers from weight degeneracy and the curse of dimensionality in high-dimensional problems (Bickel et al., 2008; Chopin, 2004; Crisan and Doucet, 2002; Gordon et al., 1993; Del Moral, 2004).

Example 2.3.2: Bootstrap particle filter in a bistable model

We return to the one-dimensional nonlinear state space model from Example 2.3.1. The extended Kalman filter struggled with the nonlinear observation function $h(x) = |x|$, and we now apply the bootstrap particle filter to the same example in order to illustrate the difference.

In Figure 2.3, we show the mean computed by the particle filter together with a confidence region in the left panel, and a subset of particle trajectories in the right panel. Here we let μ and σ denote the mean and standard deviation from the particles over time. In this particular illustration of the bootstrap particle filter, resampling is performed only when the effective sample size falls below a prescribed threshold. We do this for pedagogical reasons in order to make the resampling procedure visible. This appears as jumps in the plotted trajectories at four time points, but it should be emphasized that these are not physical jumps in the underlying particles, only the result of the resampling step.

In this case, the particle filter is more robust than the extended Kalman filter, since it can maintain a non-Gaussian and multimodal approximation and represent a large uncertainty around both wells, as the observations O do not contain information about the sign of the state S. Hence, as we see in the left panel, the mean μ of the filter does not contain much value, but we see that the the confidence (uncertainty) region well covers the true state S.

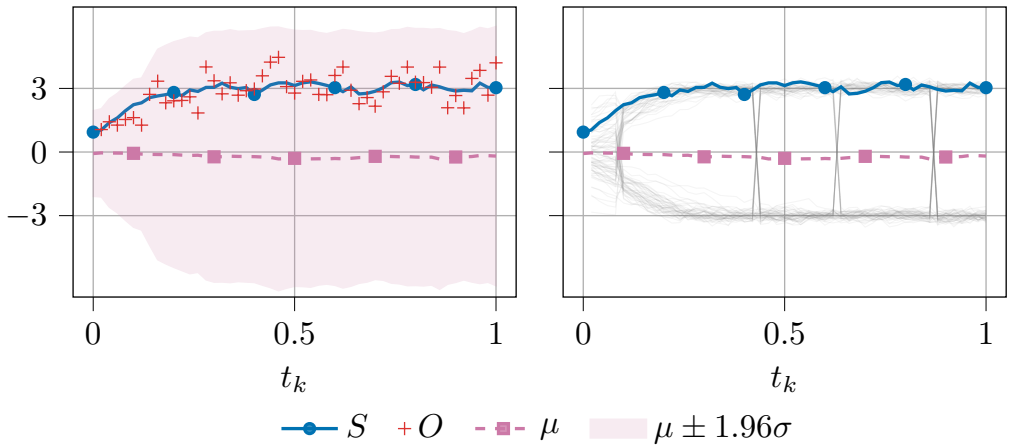


Figure 2.3: Illustration of the bootstrap particle filter for the bistable state model with observation function $h(x) = |x|$. The left panel shows the true state, the observations, and the particle filter mean with a corresponding 95% confidence region. The right panel shows a subset of particle trajectories (in gray).

The main interest of this thesis lies in the nonlinear setting, since we lack analytical solutions. Thus, the methods described above serve as central benchmarks when we develop new approximate filters in Papers I–IV.

3 Stochastic analysis and partial differential equations

This chapter highlights the most relevant mathematical background to understand both the context of the considered filtering problems and the framework from which our approximative methods are derived. In Section 3.1 we briefly outline the background from stochastic analysis in order to understand the setting, where the state S is the solution to a stochastic differential equation. In Section 3.2 we explain key concepts in partial differential equations and highlight aspects that are particularly relevant for this thesis, including the Kolmogorov equations. In Section 3.3 we connect stochastic analysis and partial differential equations through Feynman–Kac formulas. This section also contains the most relevant material for the theoretical aspects of the methodologies presented in Chapter 4. Finally, in Section 3.4 we briefly discuss approximation errors, with emphasis on time discretization errors, which play an important role in the analysis of the numerical methods considered later in the thesis.

3.1 Stochastic analysis

In this section we provide an introduction to the continuous dynamics of the signal S that we want to infer from observations. Before introducing the mathematical concepts from stochastic analysis (and later partial differential equations), we emphasize that the presentation is kept at a high level, without rigorous details from measure theory or probability theory. The main ideas should be easy to follow with a sufficient background in probability theory, and even without such a background one should still be able to gain an overview of the topic. Thus, we refrain from explicitly defining the probability space on which we work, but we use the notions of outcomes and realizations to highlight the randomness and the implicit dependence on such a space.

To this end, we introduce a stochastic differential equation whose governing coefficients, denoted by $b\colon \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$, dictate the behavior of S. The stochastic component of the equations that we study is given by a so called Brownian motion, a stochastic process that is continuous in time. For background on Brownian motion, Itô calculus, and stochastic differential equations, we refer to Øksendal (2003); Karatzas and Shreve (1991); Da Prato (2014); Friedman (1975).

Definition 3.1.1: Brownian motion

A d-dimensional *Brownian motion* $B = (B_t)_{t\geq 0}$ is a stochastic process with values in $\mathbb{R}^d$ satisfying:

1. $B_0 = 0$.
2. The sample paths $t \mapsto B_t$ are almost surely continuous.
3. $(B_t)_{t\geq 0}$ has independent increments.
4. For $0 \leq s < t$, the increment $B_t - B_s$ is normally distributed as

$$B_t - B_s \sim \mathcal{N}(0, (t-s)I_d).$$

Brownian motion serves as the fundamental source of randomness in the continuous time stochastic models that we study. Its independent and normally distributed increments make it a natural limit of random walk models and a canonical driving process for stochastic differential equations. Brownian motion exhibits highly irregular sample paths: almost surely, they are nowhere differentiable. In Figure 3.1 we see simulated realizations of Brownian motion trajectories.

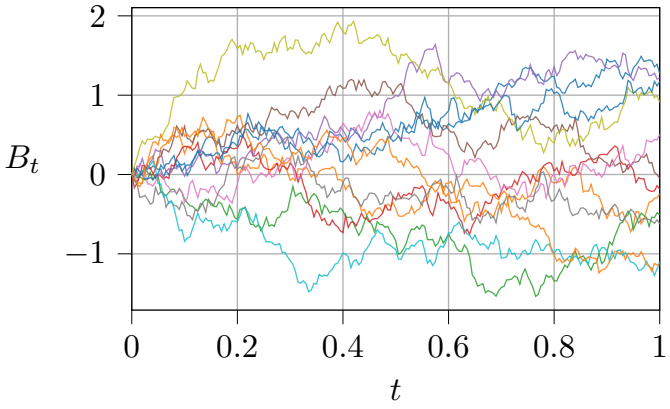


Figure 3.1: Simulated sample paths of one-dimensional Brownian motion.

The next step is to define the Stochastic Differential Equation (SDE). To this end, we assume that b, called the drift coefficient, and σ, also known as the diffusion or dispersion coefficient, are of sufficient regularity (Øksendal, 2003; Karatzas and Shreve, 1991). We say that a stochastic process $S = (S_t)_{t\in[0,T]}$ is a solution to the SDE, with drift b and diffusion σ, if it satisfies

$$S_t = S_0 + \int_0^t b(S_u)\,\mathrm{d}u + \int_0^t \sigma(S_u)\,\mathrm{d}B_u, \qquad t \in [0,T]. \tag{3.1}$$

The second integral is a so called Itô integral, where instead of integrating against the Lebesgue measure we integrate with respect to the Brownian motion B. This constitutes the randomness in the equation, and hence different realizations of S will, in general, differ, provided that σ is not identically zero, in which case the equation reduces to an ordinary differential equation. The initial value S_0 can be deterministic, $S_0 = x$ for some $x \in \mathbb{R}^d$, but often we consider it random and distributed according to some density p_0. The concept of Itô integrals, and stochastic integrals in general, is non-trivial and requires a careful study which we refrain from discussing in detail in this thesis.

In the filtering setting, the solution S to (3.1) is the underlying hidden signal that we aim to estimate from observations. In Papers I–IV we develop new methods building on stochastic analysis, and to provide background for these methods we introduce the Kolmogorov equations and the related Feynman–Kac representations. Thus, for the remainder of this chapter we focus on establishing this background rather than on the filtering problem itself.

Connected to the SDE (3.1) is a second-order differential operator A, commonly referred to as the infinitesimal generator of the SDE. With $a = \sigma\sigma^\top$, the operator is defined, for functions $\varphi \in C^2(\mathbb{R}^d;\mathbb{R})$, by

$$A\varphi(x) = \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2\varphi}{\partial x_i \partial x_j}(x) + \sum_{i=1}^{d} b_i(x)\frac{\partial\varphi}{\partial x_i}(x). \tag{3.2}$$

The infinitesimal generator plays a central role in connecting stochastic differential equations to partial differential equations (Da Prato, 2014; Friedman, 1975) and stems from Markov theory. In particular, this operator is directly related to the backward Kolmogorov equation, which we describe in Section 3.2. With this operator defined, we can recall a crucial theorem in stochastic analysis, and perhaps the most famous one, namely Itô's formula. Next we state this theorem for the case of solutions to SDEs.

Theorem 3.1.1: Itô's formula

Let $(X_t)_{t\in[0,T]}$ solve

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \qquad t \in [0,T],$$

where $(W_t)_{t\in[0,T]}$ is a Brownian motion, and X_0 is either deterministic or a random variable. If $\varphi \in C^{1,2}([0,T]\times\mathbb{R}^d;\mathbb{R})$, then, for all $0 \le r \le t \le T$, it holds that

$$\begin{aligned}\varphi(t,X_t) = \varphi(r,X_r) &+ \int_r^t \left(\frac{\partial}{\partial s} + A\right)\varphi(s,X_s)\,\mathrm{d}s \\ &+ \int_r^t \nabla\varphi(s,X_s)^\top \sigma(X_s)\,\mathrm{d}W_s.\end{aligned}$$

The theorem is widely applicable, and solutions to some of the simplest SDEs can be found by direct application of the formula, as we see next.

Example 3.1.1: Geometric Brownian motion

Consider the one-dimensional stochastic differential equation

$$X_t = X_0 + \int_0^t \alpha X_s\,\mathrm{d}s + \int_0^t \beta X_s\,\mathrm{d}W_s, \qquad X_0 > 0, \qquad t \in [0,T],$$

where $\alpha \in \mathbb{R}$ and $\beta > 0$. For $\varphi(t,x) = \log(x)$, $(t,x) \in [0,T]\times(0,\infty)$, and by inserting the coefficients $b(x) = \alpha x$ and $\sigma(x) = \beta x$ into (3.2) one obtains

$$A\varphi(t,x) = \alpha - \frac{1}{2}\beta^2, \qquad (t,x) \in [0,T]\times(0,\infty).$$

Applying Itô's formula on $[0,t]$ yields

$$\begin{aligned}\log(X_t) &= \log(X_0) + \int_0^t \left(\alpha - \frac{1}{2}\beta^2\right)\mathrm{d}s + \int_0^t \beta\,\mathrm{d}W_s \\ &= \log(X_0) + \left(\alpha - \frac{1}{2}\beta^2\right)t + \beta W_t.\end{aligned}$$

Taking the exponential on both sides gives the explicit solution

$$X_t = X_0 \exp\Big(\Big(\alpha - \frac{1}{2}\beta^2\Big)t + \beta W_t\Big), \qquad t \in [0, T].$$

The solution process X is called a geometric Brownian motion. It is a strictly positive diffusion process commonly used to model quantities exhibiting proportional growth with multiplicative noise, that is noise which scales by the process X. Some examples include asset prices in mathematical finance and stochastic population growth models (Black and Scholes, 1973; Engen, 2007). In Figure 3.2 we illustrate a few trajectories of the process with parameters $X_0 = 1$, $\alpha = 0.5$, and $\beta = 0.7$.

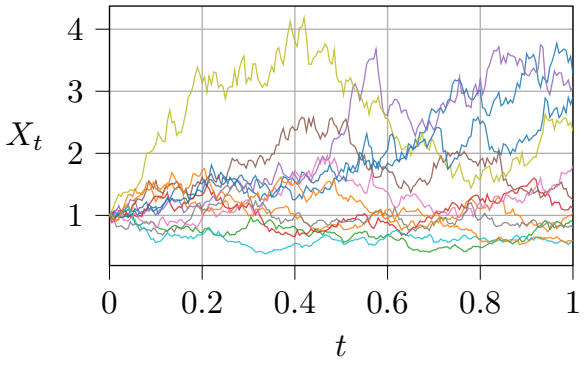


Figure 3.2: Simulated sample paths of geometric Brownian motion. The trajectories remain strictly positive and display stochastic exponential growth.

Example 3.1.2: Ornstein–Uhlenbeck process

Consider the one-dimensional stochastic differential equation

$$X_t = X_0 + \int_0^t \alpha(\theta - X_s)\,\mathrm{d}s + \int_0^t \beta\,\mathrm{d}W_s, \qquad t \in [0, T],$$

where $\alpha > 0$, $\theta \in \mathbb{R}$, and $\beta > 0$. We apply Itô's formula again and consider $\varphi(t, x) = e^{\alpha t}x$, which satisfies

$$\frac{\partial^2}{\partial x^2}\varphi(t, x) = 0.$$

Applying Itô's formula to $\varphi(t, X_t)$, we obtain

$$e^{\alpha t} X_t = X_0 + \alpha\theta \int_0^t e^{\alpha s} \, \mathrm{d}s + \beta \int_0^t e^{\alpha s} \, \mathrm{d}W_s.$$

Rearranging yields the explicit solution

$$X_t = \theta + (X_0 - \theta)e^{-\alpha t} + \beta \int_0^t e^{-\alpha(t-s)} \, \mathrm{d}W_s, \qquad t \in [0, T].$$

This solution process X is called the Ornstein–Uhlenbeck process and is used as a benchmark example in Papers I–IV. It is a mean-reverting process with long-term mean θ, which one can see in Figure 3.3 where we illustrate some trajectories of X, with $\theta = 10$, $\alpha = 1$, $\beta = 1$, and $X_0 \sim \mathcal{N}(0, 1)$. The Ornstein–Uhlenbeck process serves as a classical model for noisy relaxation phenomena in physics, short-rate dynamics in finance, continuous time Gauss–Markov modelling in signal processing and filtering, and stochastic models of neuronal activity in neuroscience (Uhlenbeck and Ornstein, 1930; Vasicek, 1977; Jazwinski, 1970; Ricciardi and Sacerdote, 1979).

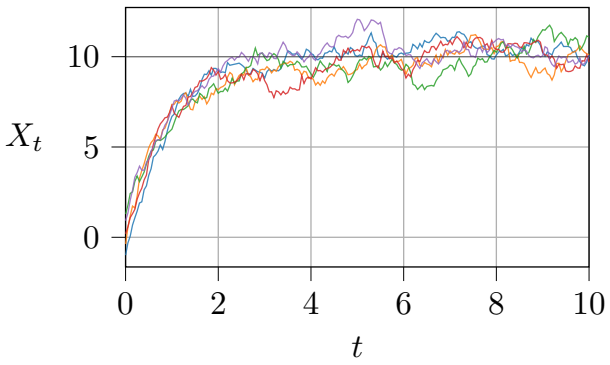


Figure 3.3: Simulated sample paths of an Ornstein–Uhlenbeck process. The trajectories fluctuate randomly while being pulled toward the long-term mean $\theta = 10$.

While analytical solutions can be derived for certain special models, such as geometric Brownian motion and the Ornstein–Uhlenbeck process, this is rarely possible in nonlinear settings. Instead, one typically needs to employ numerical approximation schemes in order to simulate trajectories of the stochastic process.

Method 3.1.1: Euler–Maruyama scheme

Consider the stochastic differential equation

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \qquad t \in [0,T]. \tag{3.3}$$

Let $0 = t_0 < t_1 < \cdots < t_N = T$ be a uniform temporal mesh with step size $\tau = T/N$. The Euler–Maruyama approximation $(\mathcal{X}_n)_{n=0}^N$ is defined recursively by

$$\mathcal{X}_{n+1} = \mathcal{X}_n + b(\mathcal{X}_n)\,\tau + \sigma(\mathcal{X}_n)\,\Delta W_n, \qquad n = 0,\ldots,N-1, \tag{3.4}$$

where the Brownian increments satisfy

$$\Delta W_n = W_{t_{n+1}} - W_{t_n} \sim \mathcal{N}(0, \tau I_d).$$

The approximation is constructed so that $\mathcal{X}_n \approx X_{t_n}$ for $n = 1,\ldots,N$ when the time step τ is sufficiently small (Kloeden and Platen, 1992).

The Euler–Maruyama scheme is the stochastic analogue of the forward Euler method for ordinary differential equations and constitutes the most commonly used time discretization for stochastic differential equations.

Example 3.1.3: Euler–Maruyama approximation of the Ornstein–Uhlenbeck process

Consider the one-dimensional Ornstein–Uhlenbeck process

$$X_t = X_0 + \int_0^t \alpha\big(\theta - X_s\big)\,\mathrm{d}s + \int_0^t \beta\,\mathrm{d}W_s, \qquad t \in [0,T],$$

Let $0 = t_0 < \cdots < t_N = T$ be a uniform time grid with step size $\tau = T/N$. Applying the Euler–Maruyama scheme yields the recursion

$$\mathcal{X}_{n+1} = \mathcal{X}_n + \alpha\big(\theta - \mathcal{X}_n\big)\tau + \beta\Delta W_n, \qquad n = 0,\ldots,N-1,$$

where the Brownian increments satisfy

$$\Delta W_n \sim \mathcal{N}(0,\tau).$$

This recursion provides a discrete time approximation of the continuous Ornstein–Uhlenbeck dynamics and converges to the exact solution (a reference solution) as the step size $\tau \to 0$, which can be seen for one trajectory in Figure 3.4. Later, in Example 3.2.3 we show that the Ornstein–Uhlenbeck process admits an explicit Gaussian distribution, and hence it serves as a natural benchmark example. In particular, moments of the approximation at given time points can be compared directly with those of the true distribution. This example is used in Papers I and IV also because, when paired with linear measurements, it leads to a linear Gaussian filtering problem for which the Kalman filter from Chapter 2 is exact.

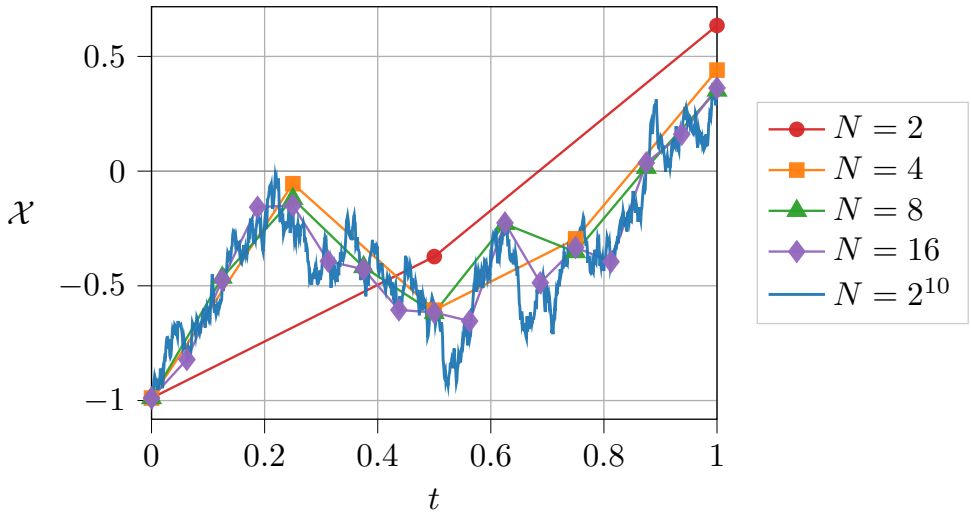


Figure 3.4: Euler–Maruyama approximations of the Ornstein–Uhlenbeck process for several step sizes, together with a highly refined reference trajectory in blue. The figure illustrates how the approximations improve as the time step decreases.

In many situations of interest, the law of S_t admits a density with respect to the Lebesgue measure, which we later denote by $p(t, \cdot)$. The existence and regularity of such densities are delicate questions and generally require structural assumptions on the coefficients of the stochastic differential equation. A powerful framework for studying these questions is provided by Malliavin calculus, which gives criteria for absolute continuity and smoothness of laws of functionals of Brownian motion; see, for example, Nualart (2006); Kusuoka and Stroock (1984); Kusuoka (2010). In particular, under nondegeneracy and, more generally, under suitable Hörmander-type conditions, one can obtain smooth transition densities for the solution process (Hairer, 2011). These ideas play

an important role in the theoretical analysis in our papers, where regularity properties of densities enter both in the formulation of the filtering problem and in the derivation of approximation results. At the same time, one should note that smoothing properties of Kolmogorov equations are not automatic outside such settings (Hairer et al., 2015).

While time-discretization schemes such as Euler–Maruyama allow us to simulate trajectories, many theoretical and numerical methods instead focus on the evolution of probability distributions and expectations associated with the process. This viewpoint leads naturally to parabolic partial differential equations.

3.2 Partial differential equations

In this section we introduce Partial Differential Equations (PDEs) that arise in close connection with stochastic differential equations. Throughout the thesis we focus exclusively on *parabolic* PDEs, which describe time-evolving diffusion phenomena and admit a direct probabilistic interpretation. These equations form the deterministic counterpart of the stochastic models introduced in Section 3.1. For classical references on parabolic PDEs and their connections to stochastic analysis, we refer to Friedman (1964, 1975); Evans (1998); Da Prato (2014); Gobet (2016).

The purpose of this section is not to give a general theory of PDEs, but rather to introduce the specific equations and interpretations that will be used later in the thesis. All discussion is therefore kept at a formal level and tailored to the filtering problem.

3.2.1 Parabolic partial differential equations

A linear second-order parabolic PDE is an equation of the form

$$\begin{aligned} \frac{\partial}{\partial t}u(t,x) &= \mathcal{L}u(t,x) + r(t,x), \qquad (t,x) \in (0,T] \times \mathbb{R}^d, \\ u(0,x) &= u_0(x), \end{aligned} \tag{3.5}$$

where $\mathcal{L}$ is a second-order differential operator acting on the spatial variable x, r is a source term, and u_0 is an initial condition. Under appropriate conditions on $\mathcal{L}$, and the functions r and u_0, there exists a (unique strong) solution u that satisfies (3.5) (Friedman, 1964; Lunardi, 1995).

Parabolic PDEs describe diffusive dynamics evolving in time and are characterized by their smoothing behavior: even if the initial condition is irregular, solutions typically become smooth for any $t > 0$.

Example 3.2.1: The heat equation

A canonical example of a parabolic PDE is the heat equation

$$\begin{aligned} \frac{\partial}{\partial t} u(t,x) &= \frac{1}{2} \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} u(t,x), && (t,x) \in (0,T] \times \mathbb{R}^d, \\ u(0,x) &= u_0(x), && x \in \mathbb{R}^d. \end{aligned}$$

The heat equation models the diffusion of temperature over time. Starting from an initial distribution u_0, the solution describes how this quantity spreads out spatially as time evolves. In Figure 3.5 we fix $d = 1$, $u_0(x) = 1_{(-\infty, 0.5]}(x)$, and $T = 0.5$, for which the exact solution u is given by

$$\begin{aligned} u(t,x) &= \Phi\left(\frac{0.5 - x}{\sqrt{t}}\right), \quad (t,x) \in (0,T] \times \mathbb{R}, \\ \Phi(z) &:= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-y^2/2} \, \mathrm{d}y, \quad z \in \mathbb{R}. \end{aligned}$$

We recall that Φ is the standard Gaussian cumulative distribution function. The smoothing behavior is easily observed in the figure, where we see the discontinuous initial condition is immediately smoothed out.

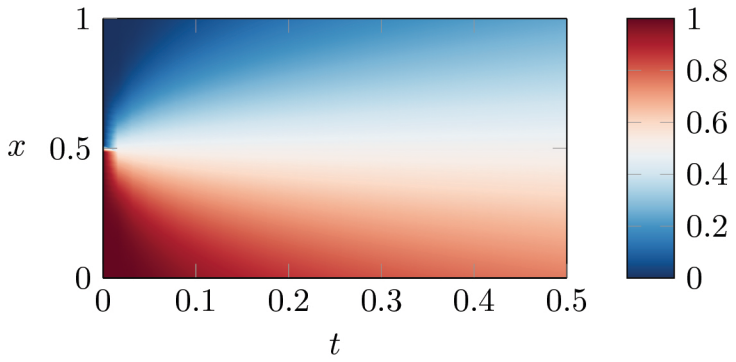


Figure 3.5: Solution of the one-dimensional heat equation with discontinuous initial condition $u_0(x) = 1_{(-\infty, 0.5]}(x)$. The figure illustrates how the initial jump is immediately smoothed out as time evolves.

From a probabilistic viewpoint, the heat equation is intimately connected to Brownian motion (Karatzas and Shreve, 1991; Øksendal, 2003). Indeed, if B is a d-dimensional Brownian motion and $u_0 \in C(\mathbb{R}^d; \mathbb{R})$, then the solution u to the heat equation can be written probabilistically as

$$u(t, x) = \mathbb{E}\big[u_0(x + B_t)\big].$$

In this sense, the deterministic spreading of heat described by the PDE corresponds to the random dispersion of Brownian trajectories.

This relationship between diffusion equations and stochastic processes serves as a simple example of a more general connection. Specifically, the connection between stochastic differential equations and parabolic PDEs is formally presented later in this chapter through the so called Feynman–Kac formula.

Parabolic PDEs of the form (3.5) arise naturally in connection with stochastic differential equations. To make this link precise, let $X = (X_t)_{t\in[0,T]}$ solve

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s,$$

where W is a d-dimensional Brownian motion. The infinitesimal generator A of X is the second-order differential operator introduced in (3.2). This operator is central in the study of stochastic differential equations. Its adjoint, denoted by A^*, is given, for functions $\varphi \in C^2(\mathbb{R}^d; \mathbb{R})$, by

$$A^*\varphi(x) = \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}(a_{ij}\varphi)(x) - \sum_{i=1}^{d} \frac{\partial}{\partial x_i}(b_i\varphi)(x). \tag{3.6}$$

These operators link the probabilistic dynamics of individual sample paths to deterministic PDEs governing expectations and probability densities. This is exactly what the so called backward and forward Kolmogorov equations model; see, for example, Friedman (1975); Da Prato (2014); Karatzas and Shreve (1991).

3.2.2 Backward Kolmogorov equation

A central concept in probability theory and stochastic analysis is the conditional expectation of a function of the terminal state. Given a test function φ, we define

$$u(t,x) := \mathbb{E}\big[\varphi(X_T) \mid X_t = x\big], \qquad (t,x) \in [0,T] \times \mathbb{R}^d. \tag{3.7}$$

Formally, this function satisfies the *backward Kolmogorov equation*

$$\begin{aligned} \frac{\partial}{\partial t} u(t,x) + Au(t,x) = 0, \qquad (t,x) \in [0,T) \times \mathbb{R}^d, \\ u(T,x) = \varphi(x). \end{aligned} \tag{3.8}$$

The equation is called backward since it is posed with a terminal condition at time T and evolves backward in time. We see how the equation equals (3.5) when $\mathcal{L} = -A$ and $r = 0$. It describes how expectations of future quantities propagate backward through the system dynamics (Friedman, 1975; Karatzas and Shreve, 1991).

Example 3.2.2: Geometric Brownian motion and an option payoff

Consider a stock price modelled by a geometric Brownian motion (Black and Scholes, 1973)

$$X_t = X_0 + \int_0^t \alpha X_s \,\mathrm{d}s + \int_0^t \beta X_s \,\mathrm{d}W_s, \qquad t \in [0,T],$$

where $\alpha \in \mathbb{R}$ is the drift and $\beta > 0$ is the volatility. Let $\varphi(x)$ be a terminal payoff function for an option depending on the stock price at final time (maturity) T. A classical example is a European call option with strike price $\kappa > 0$,

$$\varphi(x) = \max\{x - \kappa, 0\}.$$

The option price $u(t,x)$ at time t when $X_t = x$ can be written as the conditional expectation (3.7). This in turn implies that the function u solves (3.8) with the terminal condition $u(T,x) = \varphi(x)$. Thus, pricing the option can be viewed either as computing an expectation over stochastic trajectories of X, or solving a parabolic PDE with a terminal condition, as in the classical Black–Scholes framework (Black and Scholes, 1973).

One can show that the solution u is given by

$$u(t,x) = xe^{\alpha(T-t)}\Phi(d_1) - \kappa\Phi(d_2),$$

where we recall that Φ is the standard Gaussian cumulative distribution function, and

$$d_1 = \frac{\ln(x/\kappa) + \left(\alpha + \frac{1}{2}\beta^2\right)(T-t)}{\beta\sqrt{T-t}}, \qquad d_2 = d_1 - \beta\sqrt{T-t}.$$

In Figure 3.6 we illustrate the solution u on $[0,5]$ with strike price $\kappa = 2$, and SDE parameters $\alpha = 0.5$, $\beta = 0.7$, and $T = 1$ as in Example 3.1.1.

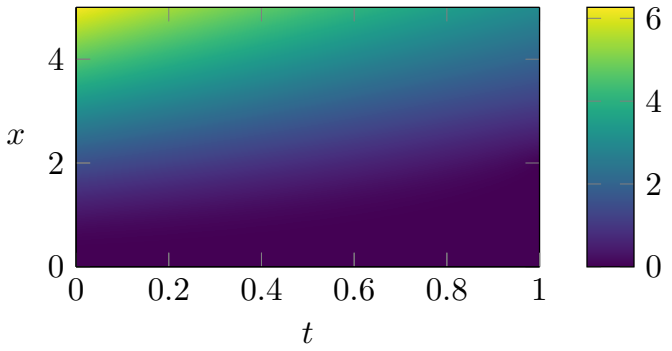


Figure 3.6: Solution of the backward Kolmogorov equation and terminal payoff $\varphi(x) = \max\{x - \kappa, 0\}$. The figure illustrates how the expected payoff depends on both time and the current state.

3.2.3 Forward Kolmogorov equation

While the backward equation governs expectations, the *forward Kolmogorov equation*, also known as the *Fokker–Planck equation*, describes the evolution of the probability density of X.

Assume the distribution of $X = (X_t)_{t\in(0,T]}$ admits a density $p = \big(p(t)\big)_{t\in(0,T]}$ and X_0 is distributed according to some density p_0. Then, formally, p satisfies

$$\frac{\partial}{\partial t}p(t,x) = A^*p(t,x), \qquad (t,x) \in (0,T] \times \mathbb{R}^d, \tag{3.9}$$

with initial condition $p(0,x) = p_0(x)$. The forward Kolmogorov equation, also known as the Fokker–Planck equation, is the PDE governing the evolution of

the density of the diffusion process (Friedman, 1975; Da Prato, 2014). In the context of filtering, it governs the prediction step between observations, where conditional densities are propagated forward in time. These prediction steps occur recursively on each time interval $[t_k, t_{k+1}]$ initiated with the filtering density at time t_k. In Chapter 4 we specify the system of equations that we aim to approximate.

Example 3.2.3: Ornstein–Uhlenbeck process and its density

Consider the one-dimensional Ornstein–Uhlenbeck process

$$X_t = X_0 + \int_0^t \alpha\big(\theta - X_s\big)\,\mathrm{d}s + \int_0^t \beta\,\mathrm{d}W_s, \qquad t \in [0,T],$$

where $\alpha > 0$, $\theta \in \mathbb{R}$, and $\beta > 0$. As shown in Example 3.1.2, this process admits the explicit representation

$$X_t = \theta + (X_0 - \theta)e^{-\alpha t} + \beta \int_0^t e^{-\alpha(t-s)}\,\mathrm{d}W_s.$$

If the initial distribution is Gaussian, that is, if $X_0 \sim \mathcal{N}(m_0, \Sigma_0)$, then X_t is Gaussian for every $t \in [0,T]$. More precisely, one can show that X_t has mean $m(t)$ and variance $\Sigma(t)$ given by

$$\begin{aligned} m(t) &= \theta + (m_0 - \theta)e^{-\alpha t}, \\ \Sigma(t) &= e^{-2\alpha t}\Sigma_0 + \frac{\beta^2}{2\alpha}\big(1 - e^{-2\alpha t}\big). \end{aligned}$$

Hence, the density of X_t is given explicitly by

$$p(t,x) = \frac{1}{\sqrt{2\pi\Sigma(t)}} \exp\!\left(-\frac{(x - m(t))^2}{2\Sigma(t)}\right), \quad (t,x) \in (0,T] \times \mathbb{R}.$$

Equivalently, the distribution of X_t can be expressed through the standard Gaussian cumulative distribution function Φ as

$$\mathbb{P}(X_t \le x) = \Phi\!\left(\frac{x - m(t)}{\sqrt{\Sigma(t)}}\right), \quad (t,x) \in (0,T] \times \mathbb{R},$$

and the density $p(t,x)$ is the corresponding derivative with respect to x. Thus, the Ornstein–Uhlenbeck process provides a canonical example in which the forward Kolmogorov equation admits an explicit Gaussian

solution. The mean is pulled towards the equilibrium level θ, while the variance converges to the limiting value $\beta^2/(2\alpha)$ as $t \to \infty$. An analogous result holds for the multivariate Ornstein–Uhlenbeck process, for which the law of X_t remains Gaussian and is fully characterized by a time-dependent mean vector and covariance matrix.

Example 3.2.4: A bistable diffusion

We consider the continuous time analogue of the bistable process in Example 2.3.1 by looking at the one-dimensional SDE

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \beta\,\mathrm{d}W_s, \qquad t \in [0,T]. \tag{3.10}$$

The drift b is induced by a potential V, that is,

$$b(x) = -\frac{\partial}{\partial x}V(x), \quad x \in \mathbb{R}, \tag{3.11}$$

where

$$V(x) = \frac{c}{4}\left(x^2 - \frac{a}{c}\right)^2, \quad x \in \mathbb{R},\ a > 0,\ c > 0. \tag{3.12}$$

This choice of drift induces two stable equilibrium points $x = \pm\sqrt{a/c}$, separated by an unstable equilibrium at $x = 0$ as shown in Figure 3.7. As a consequence, sample paths of X typically spend long periods of time fluctuating around one of the two wells, occasionally transitioning to the other due to the noise.

A single realization of (3.10) typically remains near one of the wells for a random time before making a noise-induced transition to the other. Simulating several trajectories reveals two preferred regions together with occasional switching events. In Figure 3.7, we fix $X_0 \sim \mathcal{N}(0,1)$, $a = 9$, $c = 1$, and $\beta = 1$.

Let $p(t,\cdot)$ denote the probability density of X at time t. By the forward Kolmogorov equation (3.9), p satisfies

$$\frac{\partial}{\partial t}p(t,x) = -\frac{\partial}{\partial x}\big(b(x)p(t,x)\big) + \frac{\beta^2}{2}\,\frac{\partial^2}{\partial x^2}p(t,x), \quad (t,x) \in (0,T] \times \mathbb{R}.$$

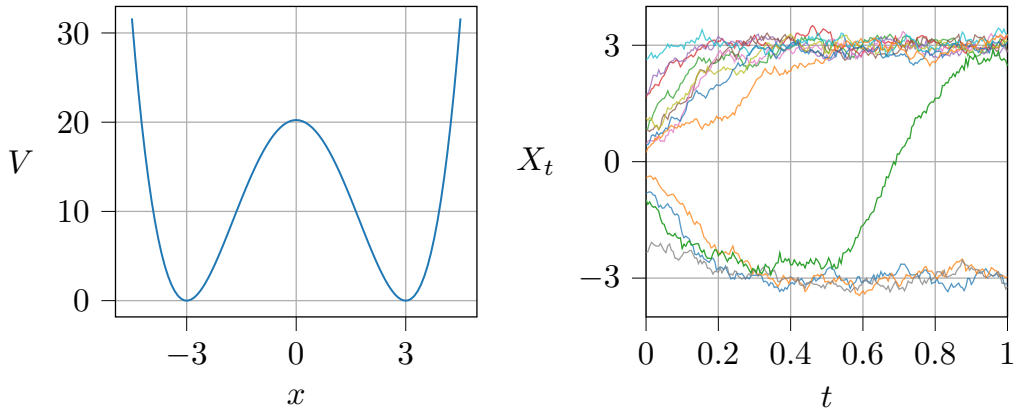


Figure 3.7: On the left, the potential V is shown, and on the right, simulated sample paths of the bistable diffusion are shown.

As t increases, the density $p(t, \cdot)$ typically becomes bimodal, reflecting the two wells of the potential in (3.12). In practice, one often solves (3.9) on a truncated spatial domain $[-D, D]$ with D sufficiently large, together with suitable boundary conditions. The resulting density can be visualized as a heat map over $[0, T] \times [-D, D]$, clearly showing the emergence of two probability peaks corresponding to the two stable states. We illustrate such a solution in Figure 3.8, approximated with a finite difference method.

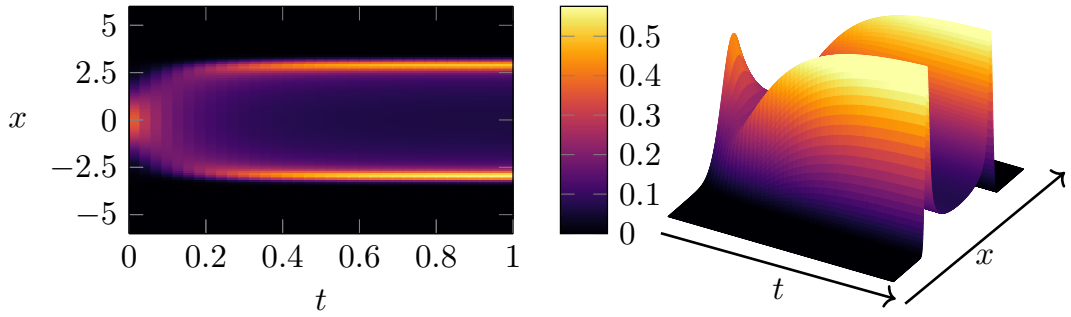


Figure 3.8: On the left, the solution to the Fokker–Planck equation is shown as a heat map. On the right, the same solution is shown as a three-dimensional surface, where the height gives the value of the density.

The PDE solution provides a deterministic, distribution-level description of the same dynamics that are observed pathwise in (3.10). Regions where $p(t, x)$ is large correspond to states where most sample paths concentrate, while low-density regions correspond to unlikely states. In particular, the bimodality of $p(t, \cdot)$ mirrors the long-term behavior of individual trajectories of X.

3.3 Feynman–Kac formulas

The Kolmogorov equations provide a deterministic description of stochastic dynamics. However, the solutions also admit probabilistic representations, as we saw in the example of the heat equation. In this section we formalize this idea through the Feynman–Kac formula, which connects linear parabolic PDEs to conditional expectations of functionals of solutions to SDEs, and nonlinear PDEs to solutions of so called backward SDEs. More precisely, we focus on Feynman–Kac formulas for the Fokker–Planck equation, directly relating to the prediction step in the filtering problem. For classical presentations of the linear Feynman–Kac formula and its PDE–probability connection, see Friedman (1975); Karatzas and Shreve (1991); Øksendal (2003).

This connection allows high-dimensional PDE problems to be treated through simulation and learning-based approximations, and forms the mathematical foundation of the methods developed in Papers I–IV.

3.3.1 Feynman–Kac representations

Assume that p is a sufficiently regular (strong) solution to (3.9). By comparing A^*, given in (3.6), with A, given in (3.2), one can identify a function $f\colon \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ such that

$$A^*\varphi(x) = A\varphi(x) + f\big(x, \varphi(x), \nabla\varphi(x)\big) \tag{3.13}$$

for sufficiently regular functions φ. Consequently, the Fokker–Planck equation (3.9) can be written in the equivalent form

$$\begin{aligned} \frac{\partial}{\partial t}p(t,x) &= Ap(t,x) + f\big(x, p(t,x), \nabla p(t,x)\big), \quad && (t,x) \in (0,T] \times \mathbb{R}^d, \\ p(0,x) &= p_0(x), && x \in \mathbb{R}^d. \end{aligned} \tag{3.14}$$

Using this form we can derive the classical Feynman–Kac representation theorem, and provide a sketch of the proof.

Theorem 3.3.1: Classical Feynman–Kac representation

Assume that there exists a unique solution $p \in C^{1,2}([0,T] \times \mathbb{R}^d; \mathbb{R})$ to (3.14) and that X is the solution to

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \qquad t \in [0,T].$$

Then, for $(t,x) \in [0,T] \times \mathbb{R}^d$, the solution p satisfies

$$p(t,x) = \mathbb{E}\Big[p_0(X_t) + \int_0^t f\big(X_s, p(t-s,X_s), \nabla p(t-s,X_s)\big)\,\mathrm{d}s \,\Big|\, X_0 = x\Big]. \tag{3.15}$$

Proof sketch of Theorem 3.3.1

We begin by defining the time-reversed function

$$v(t,x) := p(T-t,x), \qquad (t,x) \in [0,T] \times \mathbb{R}^d. \tag{3.16}$$

A direct calculation shows that v satisfies the backward parabolic PDE

$$\frac{\partial}{\partial t} v(t,x) + Av(t,x) = -f\big(x, v(t,x), \nabla v(t,x)\big), \quad (t,x) \in (0,T) \times \mathbb{R}^d, \tag{3.17}$$

with terminal condition $v(T,x) = p_0(x)$, $x \in \mathbb{R}^d$. We now apply Itô's formula to the process $t \mapsto v(t, X_t)$. For $0 \le s \le t \le T$, this gives

$$\begin{aligned} v(t,X_t) = v(s,X_s) &+ \int_s^t \Big(\frac{\partial}{\partial r} + A\Big) v(r,X_r)\,\mathrm{d}r \\ &+ \int_s^t \nabla v(r,X_r)^\top \sigma(X_r)\,\mathrm{d}W_r. \end{aligned}$$

Using (3.17), we obtain

$$\begin{aligned} v(t,X_t) = v(s,X_s) &- \int_s^t f\big(X_r, v(r,X_r), \nabla v(r,X_r)\big)\,\mathrm{d}r \\ &+ \int_s^t \nabla v(r,X_r)^\top \sigma(X_r)\,\mathrm{d}W_r. \end{aligned} \tag{3.18}$$

Letting $s = 0$ and taking expectations conditional on $X_0 = x$, and using that the Itô integral has zero conditional expectation, yields

$$v(0,x) = \mathbb{E}\Big[v(t,X_t) + \int_0^t f\big(X_s, v(s,X_s), \nabla v(s,X_s)\big)\,\mathrm{d}s \,\Big|\, X_0 = x\Big].$$

Finally, rewriting in terms of p by (3.16), we recover (3.15).

This type of probabilistic representation is classical; see, for example, Friedman (1975); Karatzas and Shreve (1991). The Feynman–Kac representation is particularly useful in high dimensions, where direct PDE solvers, such as finite element methods or finite differences (Brenner and Scott, 2008; LeVeque, 2007; Thomée, 2006, 2001; Larsson and Thomée, 2003), typically become infeasible. It allows the evaluation of $p(t,x)$ through simulation of sample paths of the SDE and the computation of corresponding path functionals. However, such approaches also usually require time discretizations, such as the Euler–Maruyama method introduced earlier in this chapter. In Chapter 4, we introduce splitting schemes with the goal of developing accurate methods for PDE that remain efficient in high dimensions, in line with our broader aim of constructing scalable Bayesian filtering methods.

The previous argument also leads naturally to a Backward Stochastic Differential Equation (BSDE) formulation. This yields a more general perspective on probabilistic representations of parabolic PDEs, and is the viewpoint underlying the deep BSDE methodology considered later in the thesis. We briefly describe this connection next through the so called nonlinear Feynman–Kac representation, which allow for possible nonlinearities in all arguments of f. For classical results on BSDEs, we refer to Pardoux and Peng (1992); El Karoui et al. (1997); Ma et al. (1994).

Theorem 3.3.2: Nonlinear Feynman–Kac representation

Assume the same setting as in Theorem 3.3.1. Then the solution p, ∇p, and X, for $t \in [0,T]$, satisfy

$$\begin{aligned} p(T-t, X_t) &= Y_t, \\ \nabla p(T-t, X_t) &= Z_t, \end{aligned}$$

where (X, Y, Z) is the solution to the forward backward stochastic

differential equation system

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \tag{3.19}$$

$$Y_t = p_0(X_T) + \int_t^T f\big(X_s, Y_s, Z_s\big)\,\mathrm{d}s - \int_t^T Z_s^\top \sigma(X_s)\,\mathrm{d}W_s. \tag{3.20}$$

Proof sketch of Theorem 3.3.2

We begin by recalling the reparameterization, $v(t) = p(T-t)$, from the proof of Theorem 3.3.1 and continue from (3.18) where we showed

$$\begin{aligned} v(t, X_t) = v(s, X_s) &- \int_s^t f\big(X_r, v(r, X_r), \nabla v(r, X_r)\big)\,\mathrm{d}r \\ &+ \int_s^t \nabla v(r, X_r)^\top \sigma(X_r)\,\mathrm{d}W_r, \end{aligned}$$

for $0 \leq s \leq t \leq T$. By defining $Y_t = v(t, X_t)$, and $Z_t = \nabla v(t, X_t)$, we get

$$Y_t = Y_s - \int_s^t f\big(X_r, Y_r, Z_r\big)\,\mathrm{d}r + \int_s^t Z_r^\top \sigma(X_r)\,\mathrm{d}W_r.$$

Choosing $s = t$ and the upper limit equal to T, then rearranging, gives

$$Y_t = Y_T + \int_t^T f\big(X_r, Y_r, Z_r\big)\,\mathrm{d}r - \int_t^T Z_r^\top \sigma(X_r)\,\mathrm{d}W_r.$$

Finally, the terminal condition in (3.19)–(3.20) follows from

$$Y_T = v(T, X_T) = p(0, X_T) = p_0(X_T).$$

The system (3.19)–(3.20) is referred to as a Forward Backward Stochastic Differential Equation (FBSDE), where the forward component is X and the backward components are (Y, Z). Since the forward equation does not depend on the backward variables, this is an uncoupled FBSDE (Ma et al., 1994; Pardoux and Peng, 1992). This is advantageous from a numerical point of view, since the approximation and simulation of X can be carried out independently of that of (Y, Z). In Chapter 4, this formulation serves as the starting point for the deep BSDE approach.

3.3.2 Outlook towards the method chapter

In Chapter 4, the Feynman–Kac viewpoint serves as a template for constructing numerical methods: one replaces the direct approximation of high-dimensional PDEs by simulation-based approximations of conditional expectations. This idea is central both for Monte Carlo-based schemes and for the learning-based approaches developed in Papers I–IV.

3.4 Approximation errors

Looking back at the Euler–Maruyama scheme introduced in Section 3.1 one might wonder how well a certain numerical scheme performs in approximating the true solution. This is not only relevant for numerical schemes for stochastic differential equations, but also for other schemes applicable to, e.g., partial differential equation. In Papers II–III we study theoretically the error of two such schemes which we develop in these papers specifically for the filtering problem.

There are several sources of approximation error relevant in this thesis, including statistical errors, optimization errors, and discretization errors. The first one can in the context of this thesis often be labeled as a Monte Carlo error, due to the common methodology of approximating expectations with averages over samples from the desired distribution. The second one is for us very much related to optimizing the parameters of a neural network, which can be seen as a high-dimensional parametric function, to minimize some objective function. The final one is discretization errors, which in this thesis consist of time discretizations, e.g., the difference between the true solution and the Euler–Maruyama scheme in Example 3.1.3.

3.4.1 Time discretization

Here, we focus on time discretization which is the error incurred due to approximating the continuous dynamics with a time discrete equivalent where we in some way or another no longer obtain the true dynamics. To give a concrete way of measuring the error we describe the (strong) error of the Euler–Maruyama method introduced in Method 3.1.1. Let X be the solution to (3.3) and $\mathcal{X}$ the approximation defined by (3.4), then the construction approximates X in the discrete time points so that $\mathcal{X}_n \approx X_{t_n}$. Following standard techniques of stochastic analysis and numerical analysis, under sufficient assumptions

on X_0, b, and σ, we can show that the difference between the true solution X and its approximation $\mathcal{X}$ goes to 0 as $N \to \infty$. We say that $\mathcal{X}$ converges in mean-square with order $\alpha > 0$ if there exist constants $C > 0$ and $N_0 \in \mathbb{N}$ such that, for all $N > N_0$,

$$\max_{n=1,\dots,N} \left(\mathbb{E}\left[\|X_{t_n} - \mathcal{X}_n\|^2 \right] \right)^{1/2} \leq C N^{-\alpha}. \tag{3.21}$$

We refer to Kloeden and Platen (1992); Gobet (2016) for rigorous presentations of the analysis.

More generally, in numerical analysis one is often interested first in determining whether a method converges to the true solution, and second in quantifying the rate at which it converges when it depends on a discretization parameter. If one can show that there exists a convergence order $\alpha > 0$ as in (3.21), then this in particular implies mean-square convergence, that is,

$$\max_{n=1,\dots,N} \left(\mathbb{E}\left[\|X_{t_n} - \mathcal{X}_n\|^2 \right] \right)^{1/2} \to 0$$

as $N \to \infty$. Returning to the earlier SDE examples, we note that the Euler–Maruyama method has convergence order $\alpha = 1$ for the Ornstein–Uhlenbeck process in Example 3.1.2 (due to the linear coefficients), while it has the standard order $\alpha = \frac{1}{2}$ for the geometric Brownian motion in Example 3.1.1 and the bistable diffusion in Example 3.2.4.

4 Density-based filtering

This chapter describes the density-based methods developed in the papers included in this thesis. To begin with, we describe in Section 4.1 the recursive system of PDEs underlying the filtering problem in the discrete-observation setting considered in Papers II–IV. We also recall the corresponding continuous-observation formulation from Paper I through the Zakai equation. The first two papers build on the idea of operator splitting, which we describe in Section 4.2. Another crucial component for all four papers is the use of optimization-based formulations for computing the probabilistic representations derived in Section 3.3. We present these in Section 4.3, and then explain in Section 4.4 how they naturally lead to regression-based approximations with neural networks.

4.1 Density-based problem formulation

We are now ready to introduce the approach to the filtering problem that is considered in Papers II–IV. Consider the state space system, with a continuous process S and a discrete observation process O given by

$$\begin{aligned} S_t &= S_0 + \int_0^t b(S_r)\,\mathrm{d}r + \int_0^t \sigma(S_r)\,\mathrm{d}B_r, \quad t \in [0,T], \\ O_k &= h(S_{t_k}) + V_k, \quad k = 1,\dots,K. \end{aligned} \tag{4.1}$$

The goal is to estimate the conditional density of S at the time points t_k given the observations $O_{1:k}$ for $k = 1,\dots,K$. If such a filtering density p_k exists, it satisfies, for a measurable set C in $\mathbb{R}^d$, the relation

$$\mathbb{P}(S_{t_k} \in C \mid O_{1:k}) = \int_C p_k(t_k, x \mid O_{1:k})\,\mathrm{d}x,$$

and more generally the conditional densities $p = (p_k)_{k=0}^K$ satisfies a system of PDE. We define the shorthand notation for the likelihood $L(o, x) = p(O_k = o \mid S_{t_k} = x)$ and let the process p be the piecewise-continuous solution representing the conditional prediction density and filter density, initialized by $p_0(0, x) = p(S_0 = x)$. The solutions are given recursively, for $k = 0, \dots, K-1$, $x \in \mathbb{R}^d$, and $o_{1:k} \in \mathbb{R}^{d' \times k}$, by

$$p_k(t, x, o_{1:k}) = p_k(t_k, x, o_{1:k}) + \int_{t_k}^{t} A^* p_k(s, x, o_{1:k}) \, \mathrm{d}s, \quad t \in [t_k, t_{k+1}], \tag{4.2}$$

$$p_{k+1}(t_{k+1}, x, o_{1:k+1}) = \frac{p_k(t_{k+1}, x, o_{1:k}) L(o_{k+1}, x)}{\int_{\mathbb{R}^d} p_k(t_{k+1}, z, o_{1:k}) L(o_{k+1}, z) \, \mathrm{d}z}. \tag{4.3}$$

The first equation is the Fokker–Planck equation (3.9) from Section 3.2, now posed on the time interval $[t_k, t_{k+1}]$ and integrated with respect to time. It is also the analogue of the prediction equation (2.1) from Chapter 2. The second part (4.3), also known as the update or correction step, is Bayes' formula and it yields an expression for the filtering density. This density-based formulation of the filtering problem, where the prediction step is represented through the Fokker–Planck equation, has also been considered in Challa and Bar-Shalom (2000); Demissie et al. (2016).

Generally, Bayesian statisticians spend a great deal of effort evaluating or approximating the denominator in the update step. This term is commonly referred to as the normalizing constant, as it normalizes the density to a probability density. In the theoretical analysis and methodology development considered in this thesis it does not matter whether we consider the normalized case, as in (4.3), or an unnormalized version given by

$$p_{k+1}(t_{k+1}, x, o_{1:k+1}) = p_k(t_{k+1}, x, o_{1:k}) L(o_{k+1}, x). \tag{4.4}$$

Using this version, however, yields an unnormalized filtering density. Utilizing this form we obtain a tractable way of evaluating the filtering density at any spatial coordinate $x \in \mathbb{R}^d$ assuming that we can find and parameterize the predictive density solution $p_k(t_{k+1})$. Hence, in the subsequent sections we fix $k = 0, \dots, K-1$ and focus on methods for approximating the prediction step (4.2). We simplify the notation and present approximation techniques for the solution p to the Fokker–Planck equation with fixed initial condition $p(0, x) = p(S_0 = x)$ on $[0, T]$, satisfying

$$p(t, x) = p(0, x) + \int_0^t A^* p(s, x) \, \mathrm{d}s, \qquad (t, x) \in (0, T] \times \mathbb{R}^d. \tag{4.5}$$

4.1.1 Continuous observation process and the Zakai equation

In Paper I, instead, we study the filtering problem when the observation process is continuous in time. We keep the notation O and let the process evolve continuously depending on S through the measurement function h, and a Brownian motion V independent of B, defined by

$$O_t = \int_0^t h(S_r)\,\mathrm{d}r + V_t. \tag{4.6}$$

For this system, (S, O), the filtering density is the solution of the Zakai equation. In the seminal paper Zakai (1969) the author derives the unnormalized filtering density $\rho(t) \propto p(S_t \mid (O_s)_{0\leq s\leq t})$ as the solution to a linear stochastic PDE. The solution ρ, for $x \in \mathbb{R}^d$ and $t \in [0, T]$, satisfies

$$\rho(t, x) = \rho(0, x) + \int_0^t A^*\rho(s, x)\,\mathrm{d}s + \int_0^t \rho(s, x)h(x)^\top\,\mathrm{d}O_s, \tag{4.7}$$

where $\rho(0, x) = p(S_0 = x)$. The form resembles the Fokker–Planck equation, where the second order operator A^* appears in the first integral, and with constant $h = 0$ one recovers the deterministic evolution equation. What differs is that we no longer have partitioned intervals, but instead integrate with respect to the observation process in the Itô integral. For classical numerical approximation of the Zakai equation, see Chow et al. (1992); Barth and Lang (2012); Frey et al. (2013). In the subsequent sections we can adapt very similar methodologies for both the Zakai equation and the Fokker–Planck equation, where the second integral of the Zakai equation becomes part of a residual first order term, which will be seen in Section 4.2.

Apart from the Zakai equation which models the unnormalized filtering density, one can also study the Kushner–Stratonovich equation modelling the normalized density (Kushner, 1964). This equation is also a stochastic PDE but no longer linear, and hence harder to study in certain regards. For both of these equations we refer to the rigorous presentations in Bain and Crisan (2009) and Xiong (2008). In Paper I we study the filtering problem through approximations of the Zakai equation.

Generally, the study of stochastic PDEs is complex and has attracted a lot of attention over the years. Especially for the case where the stochastic term (O in (4.7)) is function-valued, the interested reader is referred to the works of Prévôt and Röckner (2007); Da Prato and Zabczyk (2014); Gawarecki and Mandrekar (2011). For numerical approximations, see Lord et al. (2014).

4.2 Splitting

In this section we outline the idea of operator splitting and how we use it in Papers I–II. We begin by describing the procedure for an Ordinary Differential Equation (ODE), that is, an equation with a one-dimensional independent variable. Then we may describe it for the Zakai equation and the Fokker–Planck equation simultaneously due to their similarities.

4.2.1 ODE

We begin with the simplest setting where operator splitting is natural: an initial value problem whose vector field can be decomposed into parts that are individually easier to solve. Let $y\colon [0,T] \to \mathbb{R}^d$ solve

$$\begin{aligned} \frac{\partial}{\partial t} y(t) &= F\big(y(t)\big), \qquad t \in (0,T], \\ y(0) &= y_0 \in \mathbb{R}^d. \end{aligned} \tag{4.8}$$

Furthermore, we assume that the drift F admits a decomposition

$$F = F^{(1)} + F^{(2)},$$

so that two subproblems can be created

$$\begin{aligned} \frac{\partial}{\partial t} y^{(1)}(t) &= F^{(1)}\big(y^{(1)}(t)\big), \qquad y^{(1)}(0) = y_0, \\ \frac{\partial}{\partial t} y^{(2)}(t) &= F^{(2)}\big(y^{(2)}(t)\big), \qquad y^{(2)}(0) = y_0. \end{aligned}$$

Denote by $\Psi_t^{(i)}\colon \mathbb{R}^d \to \mathbb{R}^d$ the exact flow map of the ith subproblem so that $\Psi_t^{(i)}(y_0) = y^{(i)}(t)$, and it should be stressed that in general $y \neq y^{(1)} + y^{(2)}$. The exact solution of (4.8) is generally not available, but we can approximate it by composing the solutions to the subproblems, provided that these are either known explicitly or can be approximated accurately numerically. For a step size $\tau = \frac{T}{N}$ and grid $t_n = n\tau$ we can construct a splitting by

$$y_{n+1} = \Psi_\tau^{(2)} \circ \Psi_\tau^{(1)}(y_n), \qquad n = 0, \dots, N-1, \tag{4.9}$$

initialized by y_0, where $y_n \approx y(t_n)$. Heuristically, this splitting, known as Lie–Trotter splitting (Trotter, 1958; Hairer et al., 2006; Pazy, 1983), corresponds to evolving under $F^{(1)}$ for time τ and then under $F^{(2)}$ for time τ. In smooth

settings, the Lie–Trotter splitting is typically first-order accurate. The key point for this thesis is not the ODE accuracy per se, but the compositional principle. We will apply the same idea to the evolution equations, Fokker–Planck in (4.5) and Zakai in (4.7), where each substep corresponds to a simpler PDE or stochastic PDE that can be handled analytically or numerically.

Example 4.2.1: ODE splitting

Consider the logistic ODE

$$\frac{\partial}{\partial t} y(t) = y(t) - y(t)^2, \qquad y(0) = y_0 > 0. \tag{4.10}$$

We split the drift $F(y) = y - y^2$, into a linear growth part and a quadratic decay part,

$$F^{(1)}(y) = y,$$
$$F^{(2)}(y) = -y^2,$$

so that each subproblem is explicitly solvable.

Step 1. The solution to $\frac{\mathrm{d}}{\mathrm{d}t} y(t) = y$ with $y(0) = \eta$ is

$$\Psi_t^{(1)}(\eta) = e^t \eta.$$

Step 2. The solution to $\frac{\mathrm{d}}{\mathrm{d}t} y(t) = -y^2$ with $y(0) = \eta$ is

$$\Psi_t^{(2)}(\eta) = \frac{\eta}{1 + t\,\eta}, \qquad t \geq 0.$$

Lie–Trotter splitting. Using (4.9) we obtain, for $\tau > 0$,

$$y_{n+1} = \Psi_\tau^{(2)}\big(\Psi_\tau^{(1)}(y_n)\big) = \Psi_\tau^{(2)}\big(e^\tau y_n\big) = \frac{e^\tau y_n}{1 + \tau\, e^\tau y_n}.$$

This update preserves positivity, which is a structural property of the true solution to (4.10), whenever $y_n > 0$. In Figure 4.1 we see both flows generated by the subproblems, the exact flow of (4.10), and the resulting Lie–Trotter scheme (4.9).

If instead we reverse the order of composition, we obtain a different, but still consistent (Hairer et al., 2006), scheme,

$$\widetilde{y}_{n+1} = \Psi_\tau^{(1)}\big(\Psi_\tau^{(2)}(y_n)\big) = e^\tau \frac{y_n}{1+\tau\, y_n}.$$

Thus, the choice of splitting and composition order influences the numerical method.

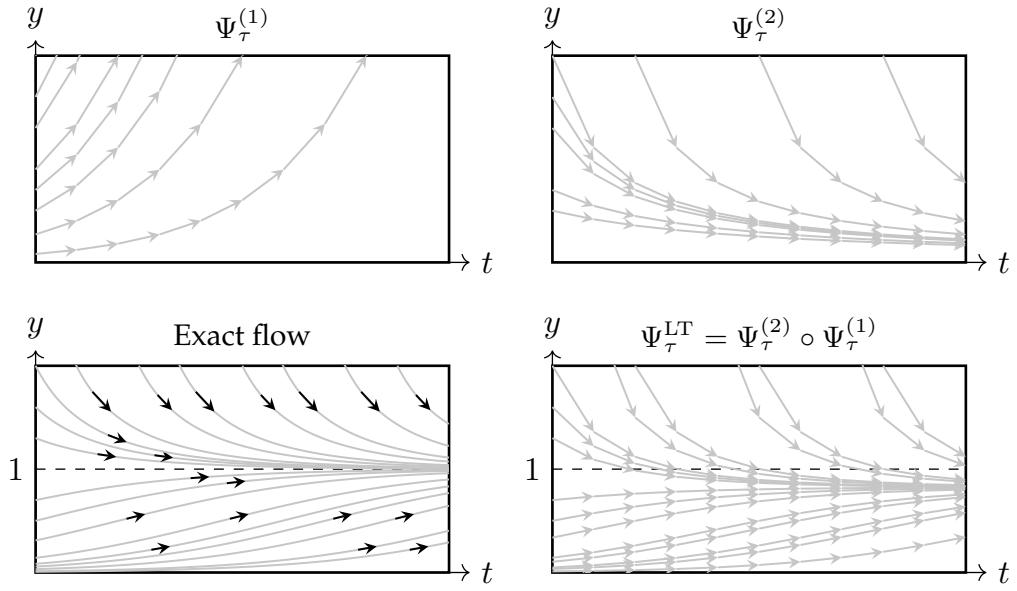


Figure 4.1: Flow portraits for the two split subproblems, the full logistic equation, and the resulting Lie–Trotter approximation with $\tau = 0.4$.

4.2.2 Filtering setting

We now apply a splitting idea to the two prediction equations, the Fokker–Planck equation (4.5) and the Zakai equation (4.7). The key observation is that both can be viewed as a dominant parabolic evolution plus a residual term. Recalling the decomposition from (3.13),

$$A^*\varphi(x) = A\varphi(x) + f\big(x, \varphi(x), \nabla\varphi(x)\big),$$

we can apply a Lie–Trotter splitting that treats the f-component first and the A-component second.

Let $\tau = \frac{T}{N}$ and $t_n = n\tau$. Suppose that π_n approximates the density at time t_n. One splitting step from t_n to t_{n+1} consists of the following two substeps, written in integral form.

Step 1: the first-order component. We first find the solution u by solving the first-order PDE, for $x \in \mathbb{R}^d$ and $t \in [t_n, t_{n+1}]$, given by

$$u(t,x) = \pi_n(x) + \int_{t_n}^{t} f\big(x, u(s,x), \nabla u(s,x)\big)\,\mathrm{d}s. \tag{4.11}$$

Step 2: the A-component. Starting from $u(t_{n+1})$, we then, for $x \in \mathbb{R}^d$ and $t \in [t_n, t_{n+1}]$, solve for v through

$$v(t,x) = u(t_{n+1},x) + \int_{t_n}^{t} Av(s,x)\,\mathrm{d}s. \tag{4.12}$$

We define $\pi_{n+1} = v(t_{n+1})$ as our approximation to $p(t_{n+1})$.

Combined splitting map. Composing the solutions of (4.11)–(4.12) recursively yields the recursion

$$\pi_n \overset{\Psi^f}{\longmapsto} u(t_{n+1}) \overset{\Psi^A}{\longmapsto} \pi_{n+1}, \qquad n = 0, \dots, N-1.$$

This is the direct analogue of Lie–Trotter splitting for ODEs from Section 4.2.1, now applied to the density evolution equation: we first evolve under the lower-order residual term and then under the parabolic diffusion operator. In practice, both substeps require numerical approximation of the PDEs.

For the Zakai equation (4.7), the same structure appears, except that the first-order step also includes the Itô integral term driven by O. Step 1 is then given by

$$u(t,x) = \pi_n(x) + \int_{t_n}^{t} f\big(x, u(s,x), \nabla u(s,x)\big)\,\mathrm{d}s + \int_{t_n}^{t} u(s,x)h(x)^\top\,\mathrm{d}O_s. \tag{4.13}$$

This is a first-order stochastic partial differential equation, and splitting-up approximations of this type have been analysed classically in the SPDE literature; in particular, Gyöngy and Krylov (2003a,b) established convergence results, including first-order convergence under suitable assumptions, for splitting schemes in time (semi-discrete) for stochastic partial differential equations. In Paper I we use this splitting technique on the Zakai equation, and in Paper II on the Fokker–Planck equation. In both cases the second split substep remains the parabolic evolution associated with A, which we focus on in the next section.

Now, with the splitting done, it remains to solve or approximate the subproblems. For Step 1, both (4.11) and (4.13) are not in general analytically solvable, and hence we employ numerical integration approximations. More precisely, we employ a forward Euler scheme (LeVeque, 2007; Thomée, 2006). For (4.11) we define the approximation U_{n+1}, for $n = 0, \dots, N-1$ and $x \in \mathbb{R}^d$, by

$$U_{n+1}(x) = \pi_n(x) + \tau f\big(x, \pi_n(x), \nabla \pi_n(x)\big).$$

Analogously, we employ the Euler–Maruyama method for the stochastic PDE (4.13) (Printems, 2001; Kruse, 2014; Lord et al., 2014), introduced for ordinary stochastic differential equations in Section 3.1. We define the approximation U_{n+1}, for $n = 0, \dots, N-1$ and $x \in \mathbb{R}^d$, by

$$U_{n+1}(x) = \pi_n(x) + \tau f\big(x, \pi_n(x), \nabla \pi_n(x)\big) + \pi_n(x) h(x)^\top \big(O_{t_{n+1}} - O_{t_n}\big).$$

Inherently the splitting itself yields only an approximation in time (semi-discrete formulation), and in the next step we see how we can reformulate the obtained scheme as an optimization problem.

4.3 Optimization-based formulations

In this section we focus on two optimization-based formulations, originating from the seminal works of Beck et al. (2021a) and E et al. (2017), respectively. The first one relates to the splitting approach above and is known as a deep splitting method, where the name originates from the combination of deep learning and splitting. The second one relates to FBSDEs which we saw an example of in Section 3.3 and is called the deep BSDE method due to its combination of deep learning and BSDEs.

4.3.1 Deep splitting approach

We outline the methodology used in Paper II, where we study the Fokker–Planck equation (4.5). Analogous explanations and derivations hold for the Zakai equation (4.7), which is treated in Paper I. We refer to Beck et al. (2021a) for the original deep splitting methodology.

In Section 3.3 we derived the classical Feynman–Kac formula for the solution p to (4.5). If we now combine the Feynman–Kac representation of p in (3.15) and the splitting with Euler approximations outlined in Section 4.2.2, we obtain an

approximation $\pi = (\pi_n)_{n=1}^N$ of p, defined recursively, for $n = 0, \dots, N-1$, by

$$\pi_{n+1}(x) = \mathbb{E}\Big[\pi_n(X_{t_{n+1}}) + \tau f\big(X_{t_{n+1}}, \pi_n(X_{t_{n+1}}), \nabla\pi_n(X_{t_{n+1}})\big) \mid X_{t_n} = x\Big].$$

Additionally, since X is generally not tractable we replace X by the Euler–Maruyama approximation $\mathcal{X}$ and define a new approximation $\overline{\pi} = (\overline{\pi}_n)_{n=1}^N$ of p, defined recursively, for $n = 0, \dots, N-1$, by

$$\overline{\pi}_{n+1}(x) = \mathbb{E}\Big[\overline{\pi}_n(\mathcal{X}_{n+1}) + \tau f\big(\mathcal{X}_{n+1}, \overline{\pi}_n(\mathcal{X}_{n+1}), \nabla\overline{\pi}_n(\mathcal{X}_{n+1})\big) \mid \mathcal{X}_n = x\Big]. \quad (4.14)$$

Now, (4.14) constitutes for each $x \in \mathbb{R}^d$ a conditional expectation that we can approximate with Monte Carlo simulations. However, due to the recursive nature we need to be able to evaluate $\overline{\pi}_n$ in the whole domain where the trajectories of $\mathcal{X}$ might end up, and doing this for every $x \in \mathbb{R}^d$ is impossible. Instead, we recast the conditional expectation as the solution to a minimization problem (see Papers I-II for details on this building on (Klenke, 2014, Corollary 8.17) and (Beck et al., 2021b, Proposition 2.7)). In the most simple form, one can think of the conditional expectation of a random variable and how it relates to mean square-minimization (where the minimization is over suitable measurable maps). More precisely, one has that for a suitable class of random variables it holds

$$Z = \mathbb{E}[X \mid Y] \iff \min_u \mathbb{E}\big[\|X - u(Y)\|^2\big] = \mathbb{E}\big[\|X - Z\|^2\big].$$

One can show that for each $n = 0, \dots, N-1$, $\overline{\pi}_{n+1}$ is the optimum to the following optimization problem

$$\min_{u \in C(\mathbb{R}^d;\mathbb{R})} \mathbb{E}\Big[\big|\overline{\pi}_n(\mathcal{X}_{n+1}) + \tau f\big(\mathcal{X}_{n+1}, \overline{\pi}_n(\mathcal{X}_{n+1}), \nabla\overline{\pi}_n(\mathcal{X}_{n+1})\big) - u(\mathcal{X}_n)\big|^2\Big]. \quad (4.15)$$

This is the basis for the optimization-formulation of the deep splitting methodology for the Fokker–Planck equation where $\overline{\pi}_n \approx p(t_n)$. In the case of continuous observations, with the Zakai equation, the objective function includes also the additional term $\overline{\pi}_n(\mathcal{X}_{n+1})h(\mathcal{X}_{n+1})^\top(O_{t_{n+1}} - O_{t_n})$ relating to the Itô integral. This optimization-based formulation is the central methodological ingredient in Papers I–II. For related developments in the direction of stochastic partial differential equations and high-dimensional nonlinear filtering, see also Beck et al. (2020); Crisan et al. (2022). These works consider settings similar to ours, but with fixed observation sequences. As a consequence, the method must be retrained for each new sequence of observations, making it less suitable for an online filtering framework.

4.3.2 Deep BSDE approach

In Paper III we develop a Bayesian filter based on the deep BSDE methodology introduced in E et al. (2017). We begin by recalling the probabilistic FBSDE formulation of the Fokker–Planck equation from Theorem 3.3.2, given by

$$\begin{aligned} X_t &= X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \\ Y_t &= p_0(X_T) + \int_t^T f(X_s, Y_s, Z_s)\,\mathrm{d}s - \int_t^T Z_s^\top \sigma(X_s)\,\mathrm{d}W_s. \end{aligned} \tag{4.16}$$

In this formulation, we recall from the derivation that $Y_t = p(T-t, X_t)$, and in particular, with $t = 0$, we have $p(T, X_0) = Y_0$. Based on this insight, and the fact that we are interested in obtaining an approximation of p directly, rather than one of Y, we reformulate the FBSDE as an optimization problem. From Theorem 3.3.2, one can show $u^*(t) = p(T-t)$, for $t \in [0,T]$, is the solution to

$$\min_{u \in C([0,T]\times\mathbb{R}^d;\mathbb{R})} \mathbb{E}\Big[\big|u(T, X_T) - p_0(X_T)\big|^2\Big] \tag{4.17}$$

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \qquad t \in [0,T], \tag{4.18}$$

$$u(t, X_t) = u(0, X_0) - \int_0^t f\big(X_s, u(s, X_s), \nabla u(s, X_s)\big)\,\mathrm{d}s \tag{4.19}$$

$$+ \int_0^t \nabla u(s, X_s)^\top \sigma(X_s)\,\mathrm{d}W_s, \quad t \in [0,T]. \tag{4.20}$$

The objective function enforces the terminal condition for Y in (4.16) in a mean-square sense. For an implementable algorithm, we discretize the stochastic processes with the Euler–Maruyama method and parameterize u by a pair $(u_0, (v_n)_{n=0}^{N-1})$, where $u_0 \approx u(0,\cdot)$ and $v_n \approx \nabla u(t_n, \cdot)$. This leads to the optimization problem

$$\min_{\substack{u_0 \in C(\mathbb{R}^d;\mathbb{R}) \\ (v_n)_{n=0}^{N-1} \in \prod_{n=0}^{N-1} C(\mathbb{R}^d;\mathbb{R}^d)}} \mathbb{E}\Big[\big|\mathcal{Y}_N - p_0(\mathcal{X}_N)\big|^2\Big] \tag{4.21}$$

$$\mathcal{X}_{n+1} = \mathcal{X}_n + \tau b(\mathcal{X}_n) + \sigma(\mathcal{X}_n)(W_{t_{n+1}} - W_{t_n}), \qquad n = 0, \dots, N-1,$$

$$\begin{aligned} \mathcal{Y}_{n+1} = u_0(\mathcal{X}_0) - \sum_{i=0}^{n} \Big(& \tau f\big(\mathcal{X}_i, \mathcal{Y}_i, v_i(\mathcal{X}_i)\big) \\ & - v_i(\mathcal{X}_i)^\top \sigma(\mathcal{X}_i)(W_{t_{i+1}} - W_{t_i}) \Big), \quad n = 0, \dots, N-1. \end{aligned}$$

Here, $\mathcal{X}_n \approx X_{t_n}$ and $\mathcal{Y}_n \approx Y_{t_n}$ denote the time-discrete approximations of the forward and backward processes. In particular, solving the minimization problem yields an approximation of the terminal density, since $u_0 \approx u(0,\cdot) = p(T,\cdot)$. For the original deep BSDE method we refer to E et al. (2017), and to Han and Long (2020); Andersson et al. (2023) for further analysis of the method, including convergence results for coupled FBSDEs. Since the resulting optimization problems are high-dimensional, we solve them using deep learning, which is introduced in the next section.

4.4 Deep learning

In Section 4.3.1–4.3.2, two optimization problems were introduced. In this section, we briefly explain how we solve them approximately. We do this through the use of neural networks, the fundamental type of models type in modern deep learning (Goodfellow et al., 2016; LeCun et al., 2015). Neural networks, in their simplest form, consist of parameterized functions Ψ^θ mapping inputs $x \in \mathbb{R}^{d_{\text{in}}}$ to outputs $y \in \mathbb{R}^{d_{\text{out}}}$, where θ denotes the trainable parameters. In this thesis, we focus on fully connected feed-forward networks, which are the primary model class used in Papers I–IV (Bishop, 1995; Goodfellow et al., 2016).

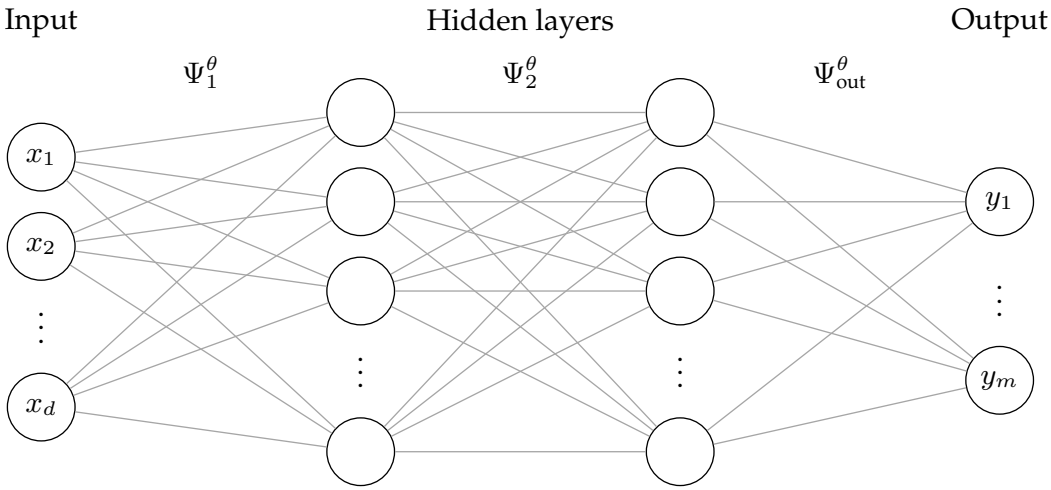
Feed-forward networks consist of L hidden layers $(\Psi_\ell^\theta)_{\ell=1}^L$, where

$$\Psi_\ell^\theta \colon \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell}, \quad \ell = 1, \dots, L,$$

together with an output layer Ψ_{out}^θ. The integers $d_0, \dots, d_L$ denote the layer widths, with $d_0 = d$ equal to the input dimension. Often, the hidden layer widths satisfy $d_\ell > d$ for some or all $\ell = 1, \dots, L$, meaning that the hidden representations are higher-dimensional than the input. The full network Ψ^θ is given by the composition

$$\Psi^\theta = \Psi_{\text{out}}^\theta \circ \Psi_L^\theta \circ \cdots \circ \Psi_1^\theta.$$

If x is the input to the ℓth layer and $y = \Psi_\ell^\theta(x)$ is the corresponding output, then full connectivity means that each component of y depends on all components of x. In the most common case, each layer consists of an affine transformation followed by a nonlinear activation function. Standard references for these constructions range from introductory expositions to more comprehensive textbook treatments (Bishop, 1995; Goodfellow et al., 2016). A schematic of a fully-connected neural network is shown in Figure 4.2.



$$y = \Psi^\theta(x) = \Psi_{\text{out}}^\theta \circ \Psi_2^\theta \circ \Psi_1^\theta(x)$$

Figure 4.2: A schematic illustration of a fully connected feed-forward neural network. Each neuron in one layer is connected to every neuron in the next layer.

In the setting of this thesis, we parameterize the continuous functions in (4.15)–(4.21) by such neural networks. Detailed specifications of the networks are given in the respective papers. In Paper I, we combine deep splitting with an energy-based approach, where the final layer is chosen with the activation function $x \mapsto e^{-x}$ so as to enforce positivity of the resulting approximation. Related energy-based approaches can be found in LeCun et al. (2006); Gustafsson et al. (2020a,b); Song and Kingma (2021). It then remains to determine the parameter values that minimize the corresponding objective functions. In practice, this is typically done using first-order gradient-based optimization methods, often in stochastic form (Goodfellow et al., 2016; Kingma and Ba, 2015; Ruder, 2016). Such methods are attractive in high-dimensional optimization because the cost of computing first-order derivatives with respect to all parameters is of the same order as the cost of evaluating the objective itself, making gradient-based training computationally feasible even for large parameter spaces (Kingma and Ba, 2015). In this thesis, optimization is carried out with standard stochastic gradient-based methods. We refer to Kingma and Ba (2015); Ruder (2016) for further discussion of practical optimization algorithms such as Adam and related variants.

5 Summary of papers

This chapter summarizes the four papers included in the thesis. The overarching goal of the papers is to develop and analyze methods that remain scalable as the state dimension of S increases. The papers follow a common structure. All of them introduce the filtering problem under consideration, develop methodology and derivations, and present numerical experiments. Paper I is different as it focuses on the filtering problem with observations that are continuous in time. Furthermore, it focuses on method development and numerical results. Papers II–III are more theoretical, with a particular focus on numerical convergence. Paper IV is primarily a numerical study emphasizing high-dimensional performance and computational efficiency.

5.1 Paper I: An energy-based deep splitting method for the nonlinear filtering problem

Summary

Paper I considers nonlinear filtering under continuous time observations. The coupled system for (S, O) is given by

$$S_t = S_0 + \int_0^t b(S_r)\,\mathrm{d}r + \int_0^t \sigma(S_r)\,\mathrm{d}B_r, \quad t \in [0,T],$$
$$O_t = \int_0^t h(S_r)\,\mathrm{d}r + V_t, \quad t \in [0,T],$$

where additional details are given in Section 4.1.1. In this setting, the filtering density is characterized by the Zakai equation (4.7). This is the equation that the paper sets out to approximate.

The proposed method combines two main ideas. The first is an *operator splitting* strategy, which decomposes the operator A^* into a second-order part equal to A A and a first-order residual term. The second is an *energy-based* regression formulation for approximating conditional densities with neural networks. Together, these yield a learning-based approximation of the filtering density. The filter is trained offline and can then be deployed online without retraining when new observations arrive.

The method is evaluated empirically on a collection of numerical examples. These include two linear equations, one of which is a high-dimensional linear spring–mass system, and two nonlinear one-dimensional equations. The paper compares the proposed method with classical baselines, including Kalman-type and particle filtering methods. The experiments indicate that the approach can produce accurate filtering estimates while also allowing fast online inference after the offline training phase.

Role in the thesis

Paper I marks the starting point of the density-based learning methodology developed in the thesis. It develops the splitting viewpoint in the continuous-observation setting through the Zakai equation and shows how energy-based learning can be used to solve the resulting filtering problem. In this way, the paper establishes several of the main ideas that reappear in later papers.

5.2 Paper II: A convergent scheme for the Bayesian filtering problem based on the Fokker–Planck equation and deep splitting

Summary

Paper II considers nonlinear filtering under discrete time observations from the system (4.1), introduced earlier in Section 4.1. We recall that S and O satisfy

$$\begin{aligned} S_t &= S_0 + \int_0^t b(S_r)\,\mathrm{d}r + \int_0^t \sigma(S_r)\,\mathrm{d}B_r, \quad t \in [0,T], \\ O_k &= h(S_{t_k}) + V_k, \quad k = 1, \dots, K. \end{aligned} \tag{5.1}$$

The filtering recursion has a prediction-update structure. The prediction step is governed by the Fokker–Planck equation, and the update is given by Bayes' formula; see (4.2)–(4.3). The main task in the paper is to approximate this recursive system. The main emphasis is on the prediction step.

The proposed method combines three ingredients. The first is an *operator splitting* strategy for the Fokker–Planck equation. The second is a probabilistic representation of the resulting subproblems through Feynman–Kac type formulas. The third is an *optimization-based* regression formulation of the corresponding conditional expectations. This makes the method amenable to approximation with neural networks and Monte Carlo sampling. As in Paper I, the resulting filter is trained offline and then applied online as new observations arrive.

A central part of the paper is a rigorous numerical analysis of the resulting deep splitting approximation. Under a parabolic Hörmander-type condition, the paper establishes a convergence rate for the approximation of the filtering density. As a corollary, it also yields convergence results for the underlying approximation of the Fokker–Planck equation itself. In simplified form, the main result shows that the approximation sequence $(\overline{\pi}_n)_{n=0}^N$ converges to the solution p of the Fokker–Planck equation at the grid points t_n, with convergence of order one. More precisely, we show

$$\sup_{x\in\mathbb{R}^d} \left|p(t_n,x) - \overline{\pi}_n(x)\right| \leq CN^{-1}, \quad n = 0,\dots,N,$$

under the assumptions stated in the paper. The paper also contains numerical experiments. Two of these verify the predicted convergence behavior: one satisfies the assumptions of the theorem, while the other does not. Even in the example where the assumptions are not satisfied, the method still appears to converge, although with a lower observed order. Finally, the paper demonstrates the method successfully on a nonlinear ten-dimensional example.

Role in the thesis

Paper II provides the first main convergence analysis in the thesis for the discrete-observation setting. It develops the deep splitting methodology into a fully analyzable filtering scheme. It also establishes a general template that reappears throughout the later papers: a prediction-update recursion, a probabilistic representation of the prediction PDE, and a regression-based optimization framework for scalable approximation. In this way, the paper forms a natural bridge between the continuous-observation perspective of Paper I and the alternative BSDE-based methodology introduced in Paper III.

5.3 Paper III: Nonlinear filtering based on density approximation and deep BSDE prediction

Summary

Paper III considers nonlinear filtering under discrete time observations (5.1). As in Paper II, the filtering problem has a prediction-update structure governed by (4.2)–(4.3). The main difficulty lies in the prediction step. In this paper, the starting point is a nonlinear Feynman–Kac representation of the prediction equation. This leads to a forward backward stochastic differential equation characterization.

The proposed method combines three main ingredients. The first is a forward backward stochastic differential equation representation of the prediction problem. The second is the deep BSDE method for approximating the backward component. The third is a density-based filtering recursion, in which the learned predictor is coupled with the Bayesian update step. This yields the deep BSDE filter. As in the previous papers, the method is trained offline and can then be deployed online without retraining during the filtering procedure.

A central part of the paper is a rigorous numerical analysis of the resulting approximation. Under a Hörmander-type condition, the paper establishes a mixed *a priori–a posteriori* error bound for the deep BSDE filter. The predicted convergence behavior is verified numerically in illustrative examples. The error between the approximation $(\overline{p}_k)_{k=1}^K$ and the true filtering density $(p_k(t_k))_{k=1}^K$ is shown to be bound by a time discretization term and a terminal residual from the BSDE approximation. More precisely, let $(\varepsilon_j)_{j=0}^{K-1}$ denote the residual values of the objective function, then the paper shows

$$\sup_{k\in\{1,\dots,K\}} \sup_{x\in\mathbb{R}^d} \left|p_k(t_k,x) - \overline{p}_k(x)\right| \leq C\left(N^{-\frac{1}{2}} + \sum_{j=0}^{K-1} \varepsilon_j\right),$$

under the assumptions stated in the paper.

Role in the thesis

Paper III contributes the second main density-based methodology in the thesis. While Paper II develops a splitting-based approach to the prediction equation, this paper shows that the same filtering objective can instead be approached

through a forward backward stochastic differential equation representation and deep BSDE approximation. Together, Papers II and III provide two complementary and analyzable routes to scalable density-based filtering. These are then compared and extended in Paper IV.

5.4 Paper IV: High-dimensional Bayesian filtering through deep density approximation

Summary

Paper IV studies nonlinear filtering in high-dimensional settings under discrete time observations. The focus is on regimes where classical particle-based methods can deteriorate because of weight degeneracy. It also considers the difficulty of directly approximating the filtering density when density values become extremely small in high dimensions.

The paper builds on the two density-based methodologies developed in Papers II and III. Both methods are extended through a logarithmic formulation. This yields positivity-preserving schemes, which is an important property of a probability density, and provides a more robust framework for higher-dimensional problems. As in the earlier papers, the methods are considered in an offline-online setting. The computationally demanding training phase is carried out offline, and the learned filters are then used for rapid online inference.

The paper is primarily empirical and comparative. It evaluates the proposed methods across a range of models and dimensions, with focus on accuracy, robustness, and computational efficiency. The numerical examples include the Ornstein–Uhlenbeck process, used as a benchmark in both high-dimensional and long-horizon settings where analytical solutions are available; the Schlögl model (Schlögl, 1972), in which the state represents a concentration undergoing chemical reactions; and the Lorenz–96 system extended to a stochastic differential equation setting in dimensions up to 100. The latter provides a demanding nonlinear high-dimensional test case and illustrates the type of scalable filtering problem targeted in the thesis.

The paper benchmarks the proposed methods against particle filters and Kalman-type baselines, including ensemble-based methods. The experiments indicate a clear distinction between low-dimensional settings, where particle methods can still perform well, and higher-dimensional settings, where the log-

arithmic density-based approaches remain more accurate and stable. The study also shows substantially faster online inference for the learned density-based methods than for particle-based alternatives.

Role in the thesis

Paper IV serves as a synthesis of the thesis and brings together the main methodological developments from the earlier papers. It places the splitting-based and BSDE-based density methods in a common benchmarking framework. It also extends both methods through positivity-preserving logarithmic formulations and demonstrates their practical scalability in higher-dimensional settings. In this way, the paper ties together the theoretical and methodological development of Papers I–III with the overall thesis objective of constructing practical and scalable methods for Bayesian filtering in high dimensions.

Bibliography

Andersson, K., Andersson, A., and Oosterlee, C. W. (2023). Convergence of a robust deep FBSDE method for stochastic control. *SIAM J. Sci. Comput.*, 45:A226–A255.

Apte, A., Jones, C. K. R. T., Stuart, A. M., and Voss, J. (2008). Data assimilation: Mathematical and statistical perspectives. *Int. J. Numer. Methods Fluids*, 56:1033–1046.

Bain, A. and Crisan, D. (2009). *Fundamentals of Stochastic Filtering*. Springer, New York.

Bao, F., Zhang, Z., and Zhang, G. (2024). A score-based filter for nonlinear data assimilation. *J. Comput. Phys.*, 514:Paper No. 113207, 16.

Bar-Shalom, Y., Li, X. R., and Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons.

Barth, A. and Lang, A. (2012). Milstein approximation for advection-diffusion equations driven by multiplicative noncontinuous martingale noises. *Appl. Math. Optim.*, 66(3):387–413.

Beck, C., Becker, S., Cheridito, P., Jentzen, A., and Neufeld, A. (2020). Deep learning based numerical approximation algorithms for stochastic partial differential equations and high-dimensional nonlinear filtering problems. *arXiv:2012.01194*.

Beck, C., Becker, S., Cheridito, P., Jentzen, A., and Neufeld, A. (2021a). Deep splitting method for parabolic PDEs. *SIAM J. Sci. Comput.*, 43:A3135–A3154.

Beck, C., Becker, S., Grohs, P., Jaafari, N., and Jentzen, A. (2021b). Solving the Kolmogorov PDE by means of deep learning. *J. Sci. Comput.*, 88(3):73.

Beneš, V. E. (1981). Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics*, 5:65–92.

Bickel, P., Li, B., and Bengtsson, T. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3, pages 318–329. Inst. Math. Statist., Beachwood, OH.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *J. Polit. Econ.*, 81(3):637–654.

Blackman, S. S. and Popoli, R. (1999). *Design and Analysis of Modern Tracking Systems*. Artech House Publishers.

Brenner, S. C. and Scott, L. R. (2008). *The Mathematical Theory of Finite Element Methods*. Springer, New York, third edition.

Brigo, D. and Hanzon, B. (1998). On some filtering problems arising in mathematical finance. *Insurance Math. Econom.*, 22(1):53–64.

Burgers, G., van Leeuwen, P. J., and Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, 126(6):1719 – 1724.

Challa, S. and Bar-Shalom, Y. (2000). Nonlinear filter design using Fokker-Planck-Kolmogorov probability density evolutions. *IEEE Trans. Aerosp. Electron. Syst.*, 36:309–315.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.*, 32(6):2385–2411.

Chopin, N., Fulop, A., Heng, J., and Thiery, A. H. (2023). Computational Doob h-transforms for online filtering of discretely observed diffusions. In *Int. Conf. Mach. Learn.*, pages 5904–5923. PMLR.

Chow, P. L., Jiang, J.-L., and Menaldi, J.-L. (1992). Pathwise convergence of approximate solutions to Zakai's equation in a bounded domain. In *Stochastic partial differential equations and applications (Trento, 1990)*, volume 268 of *Pitman Res. Notes Math. Ser.*, pages 111–123. Longman Sci. Tech., Harlow.

Crisan, D. and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Trans. Signal Process.*, 50(3):736–746.

Crisan, D., Lobbe, A., and Ortiz-Latorre, S. (2022). An application of the splitting-up method for the computation of a neural network representation for the solution for the filtering equations. *Stoch. Partial Differ. Equ.: Anal. Comput.*, 10:1050–1081.

Da Prato, G. (2014). *Introduction to Stochastic Analysis and Malliavin Calculus*. Edizioni della Normale, Pisa, third edition.

Da Prato, G. and Zabczyk, J. (2014). *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, Cambridge, second edition.

Date, P. and Ponomareva, K. (2011). Linear and non-linear filtering in mathematical finance: a review. *IMA J. Manag. Math.*, 22(3):195–211.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, New York.

Demissie, B., Khan, M. A., and Govaers, F. (2016). Nonlinear filter design using Fokker-Planck propagator in Kronecker tensor format. In *2016 19th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

E, W., Han, J., and Jentzen, A. (2017). Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Commun. Math. Stat*, 5:349–380.

El Karoui, N., Peng, S., and Quenez, M. C. (1997). Backward stochastic differential equations in finance. *Math. Finance*, 7(1):1–71.

Engen, S. (2007). Stochastic growth and extinction in a spatial geometric Brownian population model with migration and correlated noise. *Math. Biosci.*, 209(1):240–255.

Evans, L. C. (1998). *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, 99(C5):10143–10162.

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, 53(4):343–367.

Finke, A. and Thiery, A. H. (2023). Conditional sequential Monte Carlo in high dimensions. *Ann. Statist.*, 51:437–463.

Frey, R., Schmidt, T., and Xu, L. (2013). On Galerkin approximations for the Zakai equation with diffusive and point process observations. *SIAM J. Numer. Anal.*, 51:2036–2062.

Friedman, A. (1964). *Partial Differential Equations of Parabolic Type*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Friedman, A. (1975). *Stochastic differential equations and applications. Vol. 1*. Academic Press, New York-London.

Fujisaki, M., Kallianpur, G., and Kunita, H. (1972). Stochastic differential equations for the non linear filtering problem. *Osaka J. Math.*, 9(1):19–40.

Gawarecki, L. and Mandrekar, V. (2011). *Stochastic Differential Equations in Infinite Dimensions with Applications to Stochastic Partial Differential Equations*. Springer, Heidelberg.

Gobet, E. (2016). *Monte-Carlo Methods and Stochastic Processes*. CRC Press, Boca Raton, FL.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. Radar Signal Process.*, 140(2):107–113.

Gustafsson, F. K., Danelljan, M., Bhat, G., and Schön, T. B. (2020a). Energy-based models for deep probabilistic regression. In *Eur. Conf. Comput. Vis.*, pages 325–343. Springer.

Gustafsson, F. K., Danelljan, M., Timofte, R., and Schön, T. B. (2020b). How to train your energy-based model for regression. *arXiv:2005.01698*.

Gyöngy, I. and Krylov, N. (2003a). On the rate of convergence of splitting-up approximations for SPDEs. In *Stochastic inequalities and applications*, volume 56, pages 301–321. Birkhäuser, Basel.

Gyöngy, I. and Krylov, N. (2003b). On the splitting-up method and stochastic partial differential equations. *Ann. Probab.*, 31:564–591.

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric Numerical Integration*. Springer-Verlag, Berlin, second edition.

Hairer, M. (2011). On Malliavin's proof of Hörmander's theorem. *B. Sci. Math.*, 135:650–666.

Hairer, M., Hutzenthaler, M., and Jentzen, A. (2015). Loss of regularity for Kolmogorov equations. *Ann. Probab.*, 43:468–527.

Han, J. and Long, J. (2020). Convergence of the deep BSDE method for coupled FBSDEs. *Probab. Uncertain. Quant. Risk*, 5:Paper No. 5, 33.

Houtekamer, P. L. and Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.*, 126:796–811.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press, Inc., New York.

Kallianpur, G. and Striebel, C. (1968). Estimation of stochastic systems: Arbitrary system process with additive white noise observation errors. *Ann. Math. Statist.*, 39(3):785–801.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82:35–45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *J. Basic Eng.*, 83:95–108.

Kamino, K., Kadakia, N., Avgidis, F., Liu, Z.-X., Aoki, K., Shimizu, T. S., and Emonet, T. (2023). Optimal inference of molecular interaction dynamics in FRET microscopy. *Proc. Natl. Acad. Sci. U.S.A.*, 120(15).

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, second edition.

Katzfuss, M., Stroud, J. R., and Wikle, C. K. (2016). Understanding the ensemble Kalman filter. *Amer. Statist.*, 70(4):350–357.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *Proc. Int. Conf. Learn. Represent.*

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, 5(1):1–25.

Klenke, A. (2014). *Probability Theory*. Springer, London, second edition.

Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin.

Kolmogorov, A. N. (1941). Interpolation and extrapolation of stationary random sequences. *Izvestiya Akademii Nauk SSSR, Seriya Matematicheskaya*, 5:3–14.

Kruse, R. (2014). *Strong and Weak Approximation of Semilinear Stochastic Evolution Equations*. Springer, Cham.

Kushner, H. J. (1964). On the differential equations satisfied by conditional probablitity densities of Markov processes, with applications. *J. Soc. Industrial Appl. Math., Series A: Control*, 2:106–119.

Kusuoka, S. (2010). Existence of densities of solutions of stochastic differential equations by Malliavin calculus. *J. Funct. Anal.*, 258:758–784.

Kusuoka, S. and Stroock, D. (1984). Applications of the Malliavin calculus. I. In *Stochastic analysis (Katata/Kyoto, 1982)*, volume 32 of *North-Holland Math. Library*, pages 271–306. North-Holland, Amsterdam.

Larsson, S. and Thomée, V. (2003). *Partial Differential Equations with Numerical Methods*. Springer-Verlag, Berlin.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. *Predicting Structured Data*, 1.

LeVeque, R. J. (2007). *Finite Difference Methods for Ordinary and Partial Differential Equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Lord, G. J., Powell, C. E., and Shardlow, T. (2014). *An Introduction to Computational Stochastic PDEs*. Cambridge University Press, New York.

Luk, E., Bach, E., Baptista, R., and Stuart, A. (2024). Learning optimal filters using variational inference. *arXiv:2406.18066*.

Lunardi, A. (1995). *Analytic Semigroups and Optimal Regularity in Parabolic Problems*. Birkhäuser Verlag, Basel.

Ma, J., Protter, P., and Yong, J. M. (1994). Solving forward-backward stochastic differential equations explicitly—a four step scheme. *Probab. Theory Related Fields*, 98(3):339–359.

Naesseth, C. A., Lindsten, F., and Schön, T. B. (2019). High-dimensional filtering using nested sequential Monte Carlo. *IEEE Trans. Signal Process.*, 67:4177–4188.

Nualart, D. (2006). *The Malliavin Calculus and Related Topics*. Springer-Verlag, Berlin, second edition.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin Heidelberg.

Pardoux, E. and Peng, S. (1992). Backward stochastic differential equations and quasilinear parabolic partial differential equations. In Rozovskii, B. L. and Sowers, R. B., editors, *Stochastic Partial Differential Equations and Their Applications*, pages 200–217. Springer, Berlin Heidelberg.

Pazy, A. (1983). *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer-Verlag, New York.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.*, 94(446):590–599.

Prévôt, C. and Röckner, M. (2007). *A Concise Course on Stochastic Partial Differential Equations*. Springer, Berlin.

Printems, J. (2001). On the discretization in time of parabolic stochastic partial differential equations. *M2AN Math. Model. Numer. Anal.*, 35(6):1055–1078.

Ricciardi, L. M. and Sacerdote, L. (1979). The Ornstein–Uhlenbeck process as a model for neuronal activity. I. Mean and variance of the firing time. *Biol. Cybern.*, 35(1):1–9.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv:1609.04747*.

Särkkä, S. and Svensson, L. (2023). *Bayesian Filtering and Smoothing*. Cambridge University Press, Cambridge, second edition.

Schlögl, F. (1972). Chemical reaction models for non-equilibrium phase transitions. *Zeitschrift für physik*, 253(2):147–161.

Song, Y. and Kingma, D. P. (2021). How to train your energy-based models. *arXiv:2101.03288*.

Stratonovich, R. L. (1960). Conditional Markov processes. *Theory of Probability and Its Applications*, 5(2):156–178.

Swerling, P. (1959). First-order error propagation in a stagewise smoothing procedure for satellite observations. Technical Report RM-2329, RAND Corporation.

Thomée, V. (2001). From finite differences to finite elements. A short history of numerical analysis of partial differential equations. *J. Comput. Appl. Math.*, 128:1–54.

Thomée, V. (2006). *Galerkin Finite Element Methods for Parabolic Problems*. Springer-Verlag, Berlin, second edition.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

Trotter, H. F. (1958). Approximation of semi-groups of operators. *Pacific J. Math.*, 8:887–919.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the Brownian motion. *Physical Review*, 36(5):823–841.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *J. Financ. Econ.*, 5(2):177–188.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. The MIT Press.

Wonham, W. M. (1964). Some applications of stochastic differential equations to optimal nonlinear filtering. *J. Soc. Industrial Appl. Math., Series A: Control*, 2(3):347–369.

Xiong, J. (2008). *An Introduction to Stochastic Filtering Theory*. Oxford University Press, Oxford.

Zakai, M. (1969). On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11:230–243.