Analysis of discretization schemes for Fokker-Planck equations and related optimality systems

Dissertation zur Erlangung

des naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg

vorgelegt von

Masoumeh Mohammadi

aus
Tehran

Würzburg, 19.02.2015

Eingereicht am: December 2014
1. Betreuer: **Prof. Dr. Alfio Borzì**, Universität Würzburg
2. Betreuer: **Prof. Dr. Mario Annunziato**, Università degli Studi di Salerno

This thesis is dedicated to
♡ my parents ♡
for supporing me with their love and wisdom

Acknowledgements

The start point of this work has been blessed by my first German words: "Ich schwöre Treue der Verfassung des Freistaates Bayern - so wahr mir Gott helfe". Thanks to God, He never left me alone and I enjoyed doing this work with His helps. I always felt His mercy in my whole life through the very kind supports of my parents. If my parents had not encourage me to go abroad and overcome my fears, I would have never start such an exciting scientific excursion. How can I be thankful to them! Neither this work nor the greatest thing that I can do in my whole life deserves to be dedicated to them for their sincere helps. They were entirely accompanying me doing this work, although they were thousands of kilometers far from me. I have to just simply say: Thank you my dear parents. The special grace of God appeared once again in my life when Prof. Alfio Borzi gave me this opportunity to study and work under his supervision. I was always wondering how he could be such a perfect supervisor with whom I did not feel unhappy even one day during more than three years of my PhD experience! I would like to express my sincere appreciations to him for his generous guidance, encouragement and constructive criticisms, which brought to the completion of this thesis. His patience and kindness had inspired me to do my best. He was not only my supervisor, but also a very supportive father who always cared a lot about my personal problems and did not let me worry about any single issue. Dear Prof. Borzi, I am thankful to you more than words can say. I would also like to sincerely thank my co-supervisor, Prof. Mario Annunziato, for his advice and critics especially in preparation of this thesis. He kindly tried to teach me how to look deeper and more precisely. Dear Prof. Annunziato, I really appreciate your patience and valuable comments. Acknowledgments are also extended to Prof. Julien Salomon for kindly agreeing to be the co-referee of this thesis. He also helped me a lot when I stayed for a short time at the University of Dauphine in Paris. Thank you dear Prof. Salomon for all your supports and helpful collaboration and stimulating discussions. I would also like to thank Prof. Endre Süli at the University of Oxford, whom I have never met. I started the study of this work by his very well written lecture notes, and later I found him such a nice professor with whom I could easily discuss via emails. He always replied me immediately with patience and invaluable detailed answers; many thanks

to him. Extremely huge thanks also to my dear sisters and brother for their endless love and support in my whole life.

Besides the scientific contents, it would not be possible to complete my thesis without having a comfortable life in Würzburg. I remember the last day of May 2011 when I moved to Germany and met Mrs. Petra Markert-Autsch for the first time. I was immediately impressed by her superbly kind personality and generous helps, which was continued during my whole stay in Würzburg. She always raised in me the feeling of being with my sisters, and made lots of memorable moments in my life with Kristina; such a lovely girl who has inherited the kind heart of her mother. My dearest Petra, I love you! I would also like to thank Mrs. Karin Krumpholz for being so kind to me. She always surprised me with her very special presents. I am happy that I met these nice people in my life. Special appreciations are also extended to my dear friend Tanvir, who always did his best to be a real brother for me, and was caring a lot about my personal life. I truly appreciate all his sincere helps. Also sincere thanks to dear Tania, Tanvir's wife, who was also my kind sister in Würzburg and made lots of wonderful times for me with her very kind invitations. A special word of thanks goes to dear Beatrice for being such an excellent office mate, in addition to being a very good friend whom I could ask any kind of questions. I am also really grateful for being surrounded by very warmhearted friends: Suttida, Gabriele, Andreas, Martin, Juri, Danken, Stephan. Thank you so much my dear friends for making the working atmosphere so pleasant and enjoyable to me; I learned a lot from each of you. I would also like to thank Dr. Alexandra Schwartz who was not a member of our group but helped me a lot in both science and personal life. Finally, huge thanks to my dear friend Razieh for very enjoyable moments she brought in my life with her pure heart when I was writing this thesis. It was a big pleasure to have such a wonderful Iranian friend in Würzburg. I wish the best for all people who had influence in starting and completing this thesis, their names are mentioned here or I had to neglect because of the space limitation. Thank you all!

Masoumeh Mohammadi
Würzburg, December 2014

Abstract

The Fokker-Planck (FP) equation is a fundamental model in thermodynamic kinetic theories and statistical mechanics. In general, the FP equation appears in a number of different fields in natural sciences, for instance in solid-state physics, quantum optics, chemical physics, theoretical biology, and circuit theory. These equations also provide a powerful mean to define robust control strategies for random models. The FP equations are partial differential equations (PDE) describing the time evolution of the probability density function (PDF) of stochastic processes. These equations are of different types depending on the underlying stochastic process. In particular, they are parabolic PDEs for the PDF of Itō processes, and hyperbolic PDEs for piecewise deterministic processes (PDP).

A fundamental axiom of probability calculus requires that the integral of the PDF over all the allowable state space must be equal to one, for all time. Therefore, for the purpose of accurate numerical simulation, a discretized FP equation must guarantee conservativeness of the total probability. Furthermore, since the solution of the FP equation represents a probability density, any numerical scheme that approximates the FP equation is required to guarantee the positivity of the solution. In addition, an approximation scheme must be accurate and stable. For these purposes, for parabolic FP equations on bounded domains, we investigate the Chang-Cooper (CC) scheme for space discretization and first- and second-order backward time differencing. We prove that the resulting space-time discretization schemes are accurate, conditionally stable, conservative, and preserve positivity. Further, we discuss a finite difference discretization for the FP system corresponding to a PDP process in a bounded domain.

Next, we discuss FP equations in unbounded domains. In this case, finite-difference or finite-element methods cannot be applied. By employing a suitable set of basis functions, spectral methods allow to treat unbounded domains. Since FP solutions decay exponentially at infinity, we consider Hermite functions as basis functions, which are Hermite polynomials multiplied by a Gaussian. To this end, the Hermite spectral discretization is applied to two different FP equations; the parabolic PDE corresponding to Itō processes, and the system of hyperbolic PDEs corresponding to a PDP process. The resulting discretized schemes are analyzed. Stability

and spectral accuracy of the Hermite spectral discretization of the FP problems is proved. Furthermore, we investigate the conservativity of the solutions of FP equations discretized with the Hermite spectral scheme.

In the last part of this thesis, we discuss optimal control problems governed by FP equations on the characterization of their solution by optimality systems. We then investigate the Hermite spectral discretization of FP optimality systems in unbounded domains. Within the framework of Hermite discretization, we obtain sparse-band systems of ordinary differential equations. We analyze the accuracy of the discretization schemes by showing spectral convergence in approximating the state, the adjoint, and the control variables that appear in the FP optimality systems. To validate our theoretical estimates, we present results of numerical experiments.

Contents

Notation

$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
I	Identity matrix
$f(x,t)$	Probability density function
$a(x,t)$	Diffusion coefficient
$b(x,t)$	Drift coefficient
$p(x,t)$	Adjoint variable
$u(t)$	control variable
$\delta(x)$	Dirac delta function
$(\Omega, \sum, P)$	Probability space
O	Big O (Landau notation)
X(t)	A d-dimensional stochastic process with index set $[t_0, T] \subset [0, \infty)$
$\xi(t)$	A (vector-valued) white noise
$W(t)$	A (vector-valued) Wiener process
$H(x)$	Hermite polynomial
$\tilde{\mathrm{H}}(x)$	Hermite function
α	Scaling factor
w_α	Weight function
$H^r_{w_\alpha}(\mathbb{R})$	Weighted Sobolev space
$\hat{f}(t)$	Hermite coefficient
$F(x)$	Flux

Chapter 1

Introduction

1.1 Motivation

The Fokker-Planck (FP) equation is a fundamental model in statistical mechanics which governs an important class of Markov processes [91]. In general, it describes the time evolution of the probability density function of random evolutionary processes, and was used by Fokker and Planck [91] to describe the Brownian motion of a free particle (i.e., in the absence of an external force); see Figure 1.1. For an historical introduction see [38].

In the past several decades, the FP equation has been used in a number of different fields in natural sciences with a wide range of application; for instance in model systems [13, 15, 57, 77, 87, 92, 98], electron relaxation in gases [99], reactive systems [70, 85, 107], polymer dynamics [101], optical bistability [14, 17, 40], nucleation [100], dielectric relaxation [30], climate models [80], biological applications [26], astrophysical problems [88, 103, 108], economics [106], ionospheric applications [71], plasma physics [60], nuclear dynamics [1], and numerous other applications such as solid-state physics, quantum optics, chemical physics and theoretical biology [37, 90, 91]. The FP equation also appears in various types of control problems. It provides a powerful tool to define robust control strategies for stochastic models as proposed in [6, 7, 8].

Because of its wide range of application, various methods of solutions for the FP equation have been proposed in scientific literature. It includes transformation to Schrodinger equations, WKB methods, and matrix continued-fraction methods; see, e.g., [91]. Analytic solutions of the FP equations can be found in some special cases, but in general they are difficult to obtain. Therefore, numerical methods have become important in approximating the solutions of the FP equations.

Depending on the problem which is modeled by the FP equation, the solution of the FP equation may be sought in a bounded or an unbounded domain. On the other hand, the FP initial value problem differs from classical evolutionary PDEs because of the additional requirements of positivity of solution and conservativeness of total probability. The main objective of this thesis is to discuss numerical methods to

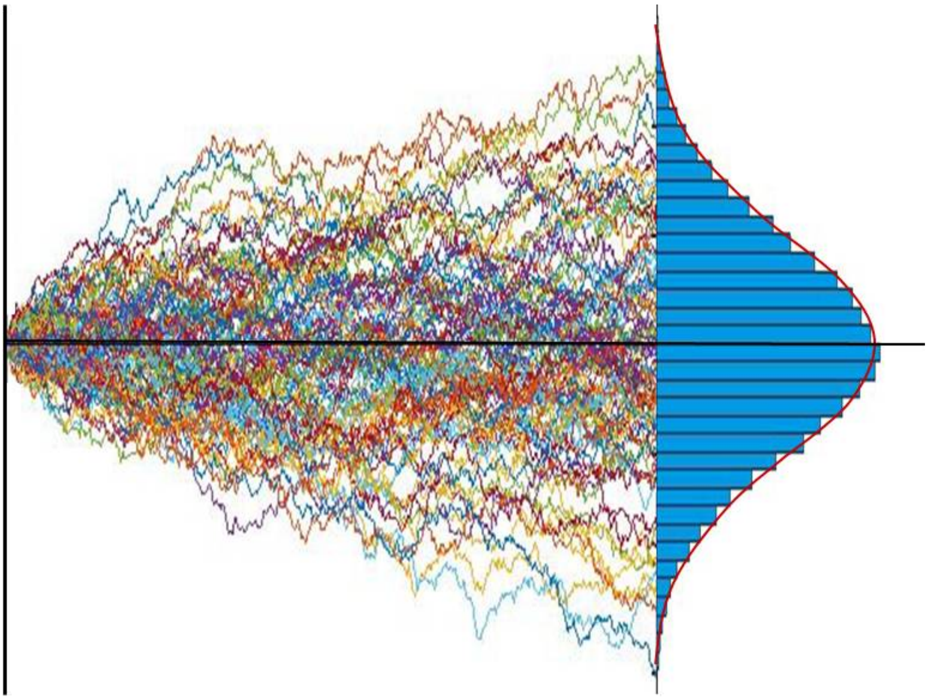


Figure 1.1: The time evolution of a stochastic process governed by a FP equation.

approximate two important classes of FP equations and related optimality systems, in both bounded and unbounded domains.

1.2 Problems and solution methods

In this thesis, we first consider the class of FP equations corresponding to Itō stochastic differential equations described by the following multidimensional model

$$\begin{cases} dX(t) = b(X(t),t)\,dt + \sigma(X(t),t)\,dW(t) \\ X(t_0) = X_0, \end{cases} \tag{1.1}$$

where the state variable $X(t) \in \mathbb{R}^d$ is subject to deterministic infinitesimal increments driven by the vector valued drift function b, and to random increments proportional to a multi-dimensional Wiener process $W(t) \in \mathbb{R}^m$, with stochastically independent components. The dispersion matrix $\sigma \in \mathbb{R}^{d\times m}$ is full rank. We notice that in most cases the state of a stochastic process can be completely characterized by the shape of its statistical distribution which is represented by the probability density function (PDF). If the initial point X_0 is a random variable which is distributed as $f^0(x)$, the evolution of the PDF associated to the stochastic process $X(t)$ is governed by the following FP model

$$\partial_t f(x,t) - \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}\left(a_{ij}(x,t)\, f(x,t)\right) + \sum_{i=1}^{d} \partial_{x_i}\left(b_i(x,t)\, f(x,t)\right) = 0, \tag{1.2}$$

$$f(x,t_0) = f^0(x), \tag{1.3}$$

defined in $Q = \Omega \times [t_0, T]$, where $\Omega \subset \mathbb{R}^d$ and f denotes the PDF function. The diffusion coefficient is given by the positive-definite symmetric matrix $a = \sigma\,\sigma^\top$, with elements

$$a_{ij} = \sum_{k=1}^{m} \sigma_{ik}\,\sigma_{jk}.$$

The initial PDF distribution f^0 must be nonnegative and normalized, $\int_\Omega f^0(x)dx = 1$. The FP model (1.2) is a parabolic problem on a multi-dimensional space domain, where the dimension corresponds to the number of components of the stochastic process. Moreover, this problem differs from a classical parabolic problem because of the additional requirements of positivity of solution and conservativeness. In fact, the FP equation guarantees the following

$$f(x,t) \geq 0, \qquad \int_\Omega f(x,t)dx = 1, \qquad \text{for all } t \geq t_0.$$

Therefore, to numerically approximate the solution of the FP model (1.2), an approximation scheme is required to be conservative and positivity-preserving in addition to be accurate and stable. For this purpose, we first focus on bounded domains and zero-flux boundary conditions and discuss finite difference discretizations. We investigate the Chang-Cooper (CC) scheme for space discretization and first- and second-order backward time differencing. Since the pioneering work of Chang and Cooper [23], different variants of this discretization strategy have been considered [21, 37, 64], that focus on first-order time discretization. From the numerical functional analytical point of view, less results are available on the accuracy and stability properties of the CC scheme. We prove that the resulting space-time discretization schemes are accurate, conditionally stable, conservative, and positivity-preserving.

We then consider another class of FP equations, where the PDF corresponds to a piecewise deterministic process (PDP). A PDP model consists of a set of differential equations that change their deterministic dynamics at random points in time. We consider a PDP model that is a first-order system of ordinary differential equations, where the driving dynamics-function is chosen by a renewal process. The d-components state function $X(t)$, $X : [t_0, \infty) \to \Omega$, $\Omega \subseteq \mathbb{R}^d$, satisfies the differential equation

$$\begin{cases} \frac{d}{dt}X(t) = A_{\mathscr{S}(t)}\left(X(t)\right), & t \in [t_0, \infty) \\ X(t_0) = X_0, \end{cases} \tag{1.4}$$

where $\mathscr{S}(t) : [t_0, \infty[\to \mathbb{S}$ is a Markov process with discrete states $\mathbb{S} = \{1, \ldots, S\}$. For each state $s \in \mathbb{S}$, we say that the dynamics is in the state s, and it is driven by the function $A_s : \Omega \to \mathbb{R}^d$, that belongs to the set of Lipschitz continuous functions $\{A_1, \dots, A_S\}$.

From a statistical point of view, the state of a PDP can be characterized by the shape of its statistical distribution which is represented by the marginal PDFs; we denote with f_s the PDF corresponding to the state s. The time evolution of these PDFs is governed by the following FP hyperbolic system [8],

$$\partial_t f_s(x,t) + \partial_x(A_s(x) f_s(x,t)) = \sum_{j=1}^{S} \mathcal{Q}_{sj} f_j(x,t), \qquad s \in \mathbb{S}, \tag{1.5}$$

where $\mathcal{Q}_{sj}$, $s, j \in \mathbb{S}$, are the components of the transition matrix $\mathcal{Q}$. Since the f_s, for $s \in \mathbb{S}$, represent the PDFs it is required that

$$\sum_{s=1}^{S} \int_{\mathbb{R}}^{d} f_s(x,t)\, dx = 1.$$

The initial conditions for the PDFs of the FP system are given as follows

$$f_s(x,0) = f_s^0(x), \qquad s \in \mathbb{S}, \tag{1.6}$$

where $f_s^0(x) \geq 0$, $x \in \mathbb{R}^d$, $\sum_{s=1}^{S} \int_{\mathbb{R}}^{d} f_s^0(x)\, dx = 1$. The model (1.5) is a first-order hyperbolic system in differential diagonal form, with coupling given through zero-order terms. This is a strictly hyperbolic model provided that the functions A_s are distinct for all $x \in \Omega$. We report theoretical results of a finite difference discretization of the FP system corresponding to a PDP with dichotomic noise [8].

Since in many cases the natural setting for stochastic processes corresponds to FP systems on unbounded domains, we discuss also the two FP models (1.2) and (1.5) in unbounded domains. In this case, finite-difference or finite-element methods cannot be applied. To treat an unbounded domain, one may truncate the domain to a bounded one and solve the problem on the bounded domain supplemented with artificial or transparent boundary conditions; see, e.g., [50, 109]. This approach is applicable for problems with rapidly decaying solutions or when exact boundary conditions are available at the truncated boundary. On the other hand, another viable approximation strategy is to consider orthogonal systems with basis functions with unbounded support. Among the orthogonal systems, Hermite functions have been used successfully in approximating the solution to parabolic FP equations in unbounded domains; see, e.g., [45, 46, 74, 75]. We further investigate the Hermite spectral approximation of hyperbolic FP models.

Next, we discuss optimal control problems governed by FP models on unbounded domains, that are investigated in our work, to find controls with the purpose of driving the PDF to attain desired objectives.

We first formulate the problem to determine a control $u \in \mathbb{R}^l$ such that starting with an initial distribution f^0 the Itō process (1.1) evolves towards a desired target probability density $f_d(x,t)$ at time $t = T$. This objective can be formulated by the following tracking functional

$$J(f,u) := \frac{1}{2}\|f(\cdot,T) - f_d(\cdot,T)\|_{w_\alpha}^2 + \frac{\nu}{2}|u|^2,$$

where $\|\cdot\|_{w_\alpha}^2$ is a weighted L^2 norm. The optimal control problem is to find u that minimizes the objective J subject to the constraint given by the FP equation (1.2). This problem can be written in a concise form as follows

$$\min_{u\in\mathbb{R}^l} J(f,u), \quad (f,u) \text{ subject to} \, (1.2) - (1.3). \tag{1.7}$$

We continue this discussion focusing on the control of a PDP FP model. We introduce in (1.5) a control mechanism in the deterministic dynamics, and consider $A_s(x,u_s)$ where $(x,u_s) \in (\Omega, U_s)$, and $U_s \subset \mathbb{R}^l$, $s \in \mathbb{S}$, are closed compact sets. We assume that for a given state (x,s) of the system, admissible open-loop control functions $u_s(t) : [0,T) \to U_s$, $s \in \mathbb{S}$, exist and are continuous in the interval $[0,T)$. Further, we assume that $A_s(x,u_s)$, $s \in \mathbb{S}$, are Lipschitz continuous and differentiable in the set (Ω, U_s) so that the differential system (1.4) has a unique solution. We consider the problem to find optimal controls u_s, $s \in \mathbb{S}$, such that the solution to the FP model (1.5) minimizes the following cost functional

$$J(f,u) := \frac{1}{2}\sum_{s=1}^{S} \|f_s(\cdot,T) - f_s^T(\cdot)\|_{w_\alpha}^2 + \frac{\nu}{2}\sum_{s=1}^{S} |u_s|_U^2,$$

where $(f_1^T, \cdots, f_S^T) \in C_0^\infty(\mathbb{R}, \mathbb{R}^S)$ is a vector of given functions with trace zero that represents a desired target PDF at time T. This objective models the requirement that the PDF of the PDP at final time, $f_s(\cdot,T)$, approaches as close as possible the desired target f_s^T. In compact form, we have the following optimal control problem

$$\min_{u\in U_1\times\ldots\times U_s} J(f,u), \quad (f,u) \text{ subject to} \, (1.5) - (1.6). \tag{1.8}$$

For the two optimal control problems (1.7) and (1.8), we derive the corresponding optimality systems and investigate their discretization in unbounded domains. We analyze the accuracy of the discretization schemes by showing spectral convergence in approximating the state, the adjoint, and the control variables.

1.3 Outline of the thesis

This thesis is organized as follows. In Chapter 2, we summarize the most important concepts on random models which are necessary to have an understanding of Itō stochastic processes and piecewise deterministic processes. In Chapter 3, we discuss the FP equations. In particular, we illustrate how to derive the FP equations of parabolic and hyperbolic type. For these models we report theoretical results concerning the existence and uniqueness of solutions. After establishing the fundamental definitions and setting for FP equations, we discretize the equations in both bounded and unbounded domains. In the case of a bounded domain, in Chapter 4 we consider finite difference discretization schemes to approximate the solution of FP equations. Specifically, for parabolic FP equations we illustrate the Chang and Cooper (CC) space-discretization scheme, and discuss this scheme in combination with a first-order backward Euler time-difference operator. We refer to this scheme as the CC-BDF scheme. We prove conditional stability and first-order in time and second-order in space accuracy. Furthermore, we show that the CC-BDF scheme is conservative and that a nonnegative initial condition results in a nonnegative solution for all times. Then, we consider a second-order backward time-differentiation formula (BDF2) and show that the resulting CC-BDF2 scheme is stable and second-order accurate in space and time and it possesses the required positivity and conservativeness properties. We discuss the CC-BDF and CC-BDF2 schemes in a one-dimensional space setting. We then discuss the extension of our results to the multi-dimensional case. At the end of Chapter 4, we present results of numerical experiments to validate the CC-BDF and CC-BDF2 schemes and our theoretical findings. In Chapter 5, we investigate the Hermite spectral discretization of the FP equations, of both parabolic and hyperbolic types, defined on unbounded domains. As a preliminary section, first the required properties and equipment for spectral methods with Hermite approximation are discussed. The accuracy of the Hermite spectral method applied to (1.2) and (1.5) is proved by showing that the error decreases spectrally as the number of expansion terms increases. Furthermore, we investigate the conservativity of the solutions of the FP equations with Hermite discretization schemes. The accuracy of the discretization method is also investigated with numerical experiments. In Chapter 6 we introduce optimal control problems. We derive the FP optimality systems consisting of state, adjoint, and optimality condition equations corresponding to the control of stochastic processes which have been studied in Chapter 2.

In Chapter 7, we present the discretization of FP optimality systems by the Hermite spectral method. The approximation method is analyzed and the accuracy of discretization schemes are discussed by showing spectral convergence in approximating the adjoint and the control variables. To further investigate the effectiveness of the method, results of numerical experiments are presented. A chapter of concluding remarks completes this thesis.

Chapter 2

Itō processes and PDP processes

The FP equations which we deal with in this thesis are modeled based on mathematical description of Itō stochastic processes and piecewise deterministic processes. We therefore summarize in this chapter the most important concepts which are necessary to have an intuitive understanding of these processes. Therefore, this chapter is of an introductory nature with the material mainly presented from [10, 47, 62, 68]. We start with recalling some basic notions and facts.

2.1 Random variables

Random variables deal with mathematical models whose outcome is determined by a random experiment. The concept of a random variable is central to this chapter.

2.1.1 Definition

Let Ω be a sample space, and let $P(\omega)$ denote the probability of an event ω. We remark that the triple $(\Omega, \sum, P)$ is called a probability space provided $\sum$ is an σ-algebra of subsets of Ω. A random variable, also known as stochastic variable, is a function X that associates a real number $X(s) = x$ with each element s of Ω. We denote by Ω_X the set of all possible values of X.

If the set Ω_X of values that the random variable X can take is finite or countably infinite, we say that X is a **discrete random variable**. A **continuous random variable** X is a random variable that can take an uncountably infinite number of values. The **distribution function** of the random variable X is defined by

$$F_X(x) = P(X \leq x) \quad \forall x \in \mathbb{R}.$$

We can define a continuous random variable as a variable whose distribution function F_X is continuous. The **probability mass function** of the discrete random variable X is defined by

$$P_X(x) = P(X = x) \quad \forall x \in \Omega_X.$$

The term (probability) *distribution* is used to designate the set of possible values of a discrete random variable, along with their respective probabilities given by the probability mass function. By extension, the same term will be employed in the *continuous* case.

2.1.2 The probability density function

The probability density function (PDF) of a continuous random variable X is defined (at all points where the derivative exists) by

$$f_X(x) = \frac{d}{dx} F_X(x). \tag{2.1}$$

Notice that the function $f_X(x)$ is not the probability $P(X = x)$ for a continuous random variable, since the probability that X is equal to some given precise value x is generally zero; that is $P(X = x) = 0$ for all $x \in \Omega_X$. The simple interpretation that can be given to $f_X(x)$ is the following:

$$\epsilon f_X(x) \simeq P(x - \frac{\epsilon}{2} \leq X \leq x + \frac{\epsilon}{2}) \tag{2.2}$$

where $\epsilon > 0$. That is, the probability that X lies between $x - \frac{\epsilon}{2}$ and $x + \frac{\epsilon}{2}$ is of the order $\epsilon f_X(x)$ for small ϵ. The equality is obtained by taking the limit as ϵ tends to zero. We have the following two essential properties for a PDF.

i) $f_X(x) \geq 0$, by the formula (2.1) and also by the formula (2.2), because F_X is a nondecreasing function.

ii) We deduce from the formula (2.1) that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

It follows that

$$\int_{-\infty}^{\infty} f_X(x) dx = 1,$$

since $F_X(\infty) = 1$.

We also have

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx.$$

Thus, the probability that X takes a value in the interval $(a, b]$ is given by the area under the curve $y = f_X(x)$ from a to b. In this thesis, we simply use the notation f instead of f_X to represent a probability density function.

2.1.3 Mean of a random variable

The mean of a discrete random variable X, also known as the expectation, mathematical expectation, expected value, or first moment, is the probability-weighted average of all possible values $x \in \Omega_X$. In other words, each possible value $x \in X$ is multiplied by its probability, and the resulting products are added together to produce the expected value. In the case of continuous random variables, the definition is the same except that the sum is replaced by an integral and the probabilities by probability densities.

More precisely, let X be a discrete random variable taking values $x \in \Omega_X$. The expected value of this random variable is the finite or infinite sum

$$E[X] = \sum_{x \in \Omega_X} x\, P(x),$$

provided that this series converges absolutely. Otherwise, we say that the expected value of X does not exist.

In the case of a continuous random variable X, we have a probability density function $f(x)$. In this case, the expected value can be computed as follows

$$E[X] = \int_{x \in \Omega_X} x f(x)\, dx.$$

The mean of a random variable X is also designated as $\langle X \rangle$, $\bar{X}$, or μ.

2.1.4 Moments and variance of a random variable

The kth **moment**, also known as the moment of order k, of the random variable X about the origin is given by $E[X^k]$, for $k = 0, 1, 2, \ldots$.

In practice, we find that the most important quantities are related to the first and second moments. In particular, for a random variable X, the variance is defined by

$$V[X] = E[(X - E[X])^2],$$

which is a nonnegative quantity. As is well known, the variance $V[X]$ or its square root the *standard deviation* $\sigma[X]$, is a measure of the degree to which the values of X deviate from the mean value X. The variance is typically designated as $Var(X)$, σ_X^2, or simply σ^2.

2.2 Normal (or Gaussian) distribution

By far the most important probability distribution is the Gaussian, or normal distribution. This is because of the central limit theorem which states that under general conditions the average of a sufficiently large number of iterates of independent random variables tends to be distributed as a normal distribution. Here we collect together the most important facts about this distribution.

The normal distribution with parameters μ and σ is a continuous distribution of a random variable X whose probability density function is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The parameter $\mu = E[X]$ in this definition is the mean or expectation of the distribution. The parameter σ is its standard deviation. Its variance is therefore σ^2, which is equal to

$$\mathrm{V}(X) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \sigma^2.$$

A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

The normal distribution is also often denoted by $\mathcal{N}(\mu, \sigma^2)$. Thus when a random variable X is distributed normally with mean μ and variance σ^2, we write

$$X \;\sim\; \mathcal{N}(\mu,\, \sigma^2).$$

2.3 Stochastic process

Suppose that with each element s of a sample space Ω of some random experiment, we associate a function $X(t, s)$, where t belongs to $\mathbb{T} \subset \mathbb{R}$. The set $\{X(t, s), t \in \mathbb{T}\}$ is called a stochastic (or random) process. The function $X(t, s)$ is a random variable for any particular value of $t \in \mathbb{T}$. The set $\mathbb{T}$ is usually the set $\mathbb{N}_0 = \{0, 1, \ldots\}$ or the interval $[0, \infty)$. For each point $s \in \Omega$, the mapping $t \mapsto X(t, s)$ is the corresponding *sample path.*

We consider the case when $\mathbb{T}$ is either a countably infinite set or an uncountably infinit set. Moreover, the set of possible values of the random variables $X(t, s)$ can be discrete (that is, finite or countably infinite) or continuous (that is, uncountably infinite). Consequently, there are four different types of stochastic processes.

If $\mathbb{T}$ is a countably infinite set (respectively, an interval or a set of intervals), then $\{X(t, s), t \in \mathbb{T}\}$ is said to be a **discrete-time** (respectively, **continuous-time**) stochastic process.

The set $\Omega_{X(t)}$ of values that the random values $X(t)$ can take is called the **state space** of the stochastic process $\{X(t, s), t \in \mathbb{T}\}$. If $\Omega_{X(t)}$ is finite or countably infinite (respectively, uncountably infinite), $\{X(t, s), t \in \mathbb{T}\}$ is said to be a **discrete-state** (respectively, **continuous-state**) process.

Since it will not be necessary to write explicitly the argument s of the function $X(t, s)$, the stochastic process will be denoted by $\{X(t), t \in \mathbb{T}\}$. However, in the discrete case, it is customary to write $\{X_n, n \in \mathbb{T}\}$.

Furthermore, corresponding to any continuous-time and continuous-state stochastic process $\{X(t), t \in \mathbb{T}\}$ we have the following two important quantities.

i) The **infinitesimal mean** $m(x;t)$ of $X(t)$ is defined by

$$m(x;t) = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} E[X(t+\epsilon) - X(t) | X(t) = x].$$

We can also obtain $m(x_0;t_0)$ as follows:

$$m(x_0;t_0) = \lim_{t \downarrow t_0} \frac{\partial}{\partial t} E[X(t) | X(t_0) = x_0]. \tag{2.3}$$

ii) The **infinitesimal variance** $v(x;t)$ of $X(t)$ is defined by

$$v(x;t) = \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} E[(X(t+\epsilon) - X(t))^2 | X(t) = x].$$

We can also obtain $v(x_0;t_0)$ as follows:

$$v(x_0;t_0) = \lim_{t \downarrow t_0} \frac{\partial}{\partial t} V[X(t) | X(t_0) = x_0]. \tag{2.4}$$

Suppose that the process $\{X(t), t \in \mathbb{T}\}$ has infinitesimal moments $m(x;t) = m(x)$ and $v(x;t) = v(x)$ for all $\mathbb{T}$, and that its state space is the interval $[a,b]$ (or $[a,b)$, etc.). Let

$$Y(t) := g(X(t)) \quad \text{for } t \in \mathbb{T}.$$

If the function g is strictly increasing or decreasing on the interval $[a,b]$ and if the second derivative $g''(x)$ exists and is continuous, for $a < x < b$, then we can show that the infinitesimal moments of the process $\{Y(t), t \in \mathbb{T}\}$ are given by

$$m_Y(y) = m(x)g'(x) + \frac{1}{2}v(x)g''(x),$$

and

$$v_Y(y) = v(x)(g'(x))^2$$

where $x = g^{-1}(y)$. Moreover, the state space of the process is the interval $[g(a), g(b)]$ (respectively, $[g(b), g(a)]$) if g is strictly increasing (respectively, decreasing).
It can be shown that the function $P(x,t,x_0,t_0)$ satisfies the following partial differential equations:

$$\partial_t P + \partial_x(m(x;t)P) - \frac{1}{2}\partial_{xx}(v(x;t)P) = 0$$

and

$$\partial_{t_0} P + m(x_0;t_0)\partial_{x_0} P + \frac{1}{2}v(x_0;t_0)\partial_{x_0x_0} P = 0.$$

These equations are called the *Kolmogorov equations.* The first one is the Kolmogorov *forward equation* (*Fokker-Planck equation*), and the second one is the Kolmogorov *backward equation.*

The Kolmogorov forward equation is used when we have information about the state x of the system at time t, and we want to know the probability distribution of the state at a later time $s > t$. The word "forward" refers to the fact that the PDE is integrated forward in time. The Kolmogorov backward equation on the other hand is useful when we want to know for every state x at time t, $(t < s)$ what is the probability of ending up in the target set at time s. In this case we integrate "backward" in time from s to t.

2.4 Wiener process

A stochastic process $\{W(t), t \geq 0\}$ is called a Wiener process if

- $W(0) = 0$
- $\{W(t), t \geq 0\}$ has independent and stationary increments,
- $W(t) \sim \mathcal{N}(0, \sigma^2 t) \quad \forall t > 0.$

The condition that it has independent increments means that if $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2$ then $W(t_1) - W(s_1)$ and $W(t_2) - W(s_2)$ are independent random variables.

From the point of view of this thesis, the probability density function of the Wiener process is the solution of the Fokker-Planck equation in which $W(t)$ is the variable of the equation, the drift coefficient is zero and the diffusion coefficient is one. Consider the partial differential equation

$$\partial_t f(w, t|w_0, t_0) = \frac{1}{2}\partial_{ww} f(w, t|w_0, t_0). \tag{2.5}$$

The solution to (2.5) is as follows [47],

$$f(w, t|w_0, t_0) = [2\pi(t - t_0)]^{-1/2} \exp[-(w - w_0)^2/2(t - t_0)].$$

This represents a Gaussian, with

$$\mu = E[W(t)] = w_0,$$

$$\sigma^2 = E[(W(t) - w_0)^2] = t - t_0,$$

so that an initially sharp distribution spreads in time.

The one-variable Wiener process is often simply called Brownian motion, since the Wiener process equation (2.5) is exactly the same as the differential equation of diffusion, shown by Einstein to be obeyed by Brownian motion.

2.5 Brownian motion

The observation that, when suspended in water, small pollen grains are found to be in a very animated and irregular state of motion, was first systematically investigated by Robert Brown in 1827, and the observed phenomenon took the name Brownian motion because of his fundamental pioneering work. Brownian motion is among the simplest of the continuous-time stochastic processes, which is described by the Wiener process.

2.6 Markov process

The ground work for the theory of Markov stochastic processes was laid in 1906 by A.A. Markov who formulated the principle that the *future* is independent of the *past* when we know the *present*. On the other hand, this principle is the causality principle of classical physics carried over to stochastic dynamic systems. It specifies that knowledge of the state of a system at a given time is sufficient to determine its state at any future time. If we carry this idea over to stochastic dynamic systems, we get the Markov property. It says that *if the state of a system at a particular time s (the present) is known, additional information regarding the behavior of the system at times $t < s$ (the past) has no effect on our knowledge of the probable development of the system at $t > s$ (the future).*

2.7 Diffusion Process

Diffusion processes are special cases of Markov processes with continuous sample functions which serve as probability models of physical diffusion phenomena. The simplest and oldest example is the motion of very small particles, such as grains of pollen in a fluid, the so-called Brownian motion.

Mathematically, a Markov process $\{X(t), t \in [t_0, T]\}$ with values in $\mathbb{R}^d$, $d \geq 1$, and almost certainly continuous sample functions is called a **diffusion process** if its transition probability P satisfies the following three conditions for every $t \in [t_0, T)$, $x \in \mathbb{R}^d$, and $\epsilon > 0$:

a)

$$\lim_{t \downarrow s} \frac{1}{t-s} \int_{|y-x|>\epsilon} P(x,t,dy,s) = 0;$$

b) there exists an $\mathbb{R}^d$-valued function $b(x,t)$ such that

$$\lim_{t \downarrow s} \frac{1}{t-s} \int_{|y-x|\leq\epsilon} (y-x) P(x,t,dy,s) = b(x,t);$$

c) there exists a $d \times d$ matrix-valued function $a(x,t)$ such that

$$\lim_{t \downarrow s} \frac{1}{t-s} \int_{|y-x|\leq\epsilon} (y-x)(y-x)^T P(x,t,dy,s) = a(x,t).$$

The functions b and a are called the coefficients of the diffusion process. In particular, b is called the **drift vector** and a is called the **diffusion matrix**. $a(x,t) \in \mathbb{R}^{d\times d}$ is symmetric and nonnegative-definite.

The decisive property of diffusion process is that their transition probability P is, under certain regularity assumptions, uniquely determined merely by the drift vector and the diffusion matrix. To each diffusion process with coefficients b and $a = (a_{ij})$ is assigned the second-order differential operator

$$\mathcal{D} \equiv \sum_{i=1}^{d} b_i(x,t)\partial_{x_i} + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}(x,t)\partial_{x_i x_j}.$$

$\mathcal{D}g$ can be formally written for every twice partially differentiable function $g(x)$ and is determined by b and a.

Diffusion process is also a generalization of the Wiener process. Let

$$Y(t) := \sigma W(t) + \mu t$$

where $\{W(t), t \geq 0\}$ is a Wiener process, and μ and $\sigma \neq 0$ are real constants. We have

$$E[Y(t)|Y(t_0) = y_0] = y_0 + \mu(t - t_0),$$

and

$$V[Y(t)|Y(t_0) = y_0] = \sigma^2(t - t_0),$$

for all $t \geq t_0$. We then deduce from the formulas (2.3) and (2.4) that

$$m_Y(y) = \mu \qquad \text{and} \qquad v_Y(y) = \sigma^2 \quad \forall y.$$

The stochastic process $\{Y(t), t \geq 0\}$ is called a diffusion process whose infinitesimal parameters are given by $m_Y(y) \equiv \mu$ and $v_Y(y) \equiv \sigma^2$. The process $\{Y(t), t \geq 0\}$ is also called a Brownian motion (or Wiener process) with drift μ.

Remark. The parameter μ is the drift coefficient, and σ^2 is the diffusion coefficient. The term *parameter*, rather than *coefficient*, is used as well.

2.8 Stochastic differential equations

A deterministic system is a system in which no randomness is involved in the development of the system. A deterministic model will thus always produce the same output from a given starting condition or initial state. In many applications, however, the system modeled by differential equations do not behave as predicted.

A stochastic differential equation (SDE) is a model in which one or more of the terms are stochastic processes, resulting in a solution which is itself a stochastic process. In many applications such equations result from the incorporation of random fluctuations in the dynamical description of a system. An example is the molecular bombardment of a speck of dust on a water surface, which results in Brownian motion [61]. The intensity of this bombardment does not depend on the state variables, for instance the position and velocity of the speck. Taking $X(t)$ as one of the components of the velocity of the particle, Langevin wrote the equation

$$\frac{d}{dt}X(t) = -\gamma X(t) + \sigma\xi(t), \tag{2.6}$$

for the acceleration of the particle. Here $\xi(t)$ is a white noise process and σ is intensity which is independent of the velocity. The *Langevin equation* (2.6) is symbolically interpreted as a stochastic differential equation

$$dX(t) = -\gamma X(t)\,dt + \sigma\,dW(t),$$

that is as a stochastic integral equation

$$X(t) = X(t_0) - \int_{t_0}^{t} \gamma X(s)\,ds + \int_{t_0}^{t} \sigma\,dW(s)$$

where the second integral is an Itō stochastic integral. This example is one of the simplest and the oldest stochastic differential equations.

2.9 Itō processes

A very wide class of continuous stochastic processes can be obtained by modeling various diffusion processes. They are generally characterized by being Markov processes and having local drift and diffusion; that is, behaving near a point x on the time interval Δt like $\sigma(x)\Delta w_t + b(x)\Delta t$, where $\sigma(x)$ is the local diffusion coefficient and $b(x)$ is the local drift. A quite satisfactory model of such processes is given by solutions of stochastic Itō equations. A stochastic quantity $X(t)$ obeys an Itō stochastic differential equation written as

$$dX(t) = b(X(t),t)\,dt + \sigma(X(t),t)\,dW(t) \tag{2.7}$$

if for all t and t_0,

$$X(t) = X(t_0) + \int_{t_0}^{t} b(X(s),s)\,ds + \int_{t_0}^{t} \sigma(X(s),s)\,dW(s).$$

The conditions which are required for existence and uniqueness of solutions to the SDE (2.7) in a time interval $[t_0, T]$ are:
i) *Lipschitz condition*: a K exists such that

$$|\sigma(x,t) - \sigma(y,t)| + |b(x,t) - b(y,t)| \leq K|x-y|$$

for all x, y and t in the range $[t_0, T]$.
ii) *growth condition*: a K exists such that for all t in the range $[t_0, T]$,

$$|\sigma(x,t)|^2 + |b(x,t)|^2 \leq K^2(1+|x|^2).$$

Under these conditions there will be a unique solution $X(t)$ in the range $[t_0, T]$. Almost every stochastic differential equation encountered in practice satisfies the Lipschitz condition since it is essentially a smoothness condition. However, the growth condition is often violated. This does not mean that no solution exists; rather, it means the solution may explode to infinity, that is, the value of x can become infinite in a finite time (blow up); in practice, a finite random time. This phenomenon occurs in ordinary differential equations. Failing to satisfy the Lipschitz condition does not necessarily imply that the solution will explode.

If b and σ are linear, that is $b(X(t),t) = b_1(t)X(t) + b_2(t)$ and $\sigma(X(t),t) = \sigma_1(t)X(t) + \sigma_2(t)$, where the coefficients b_1, b_2, σ_1, σ_2 are specified functions of time, we have a linear Itō stochastic differential equation. When $\sigma_1(t) \equiv 0$ we say that the SDE is *linear in the narrow sense.*

2.10 The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck process is the historically oldest example of a stochastic differential equation. As already remarked in subsection 2.8, for the Brownian motion of a particle under the influence of friction we have the Langevin equation. The corresponding stochastic differential equation

$$dX(t) = -\gamma X(t)\,dt + \sigma\,dW(t), \quad X(0) = c,$$

is linear in the narrow sense and autonomous. Therefore, its unique solution is given by

$$X(t) = e^{-\alpha t}c + \sigma \int_0^t e^{-\gamma(t-s)}dW(s),$$

where the integration is the Itō stochastic integral defined, e.g., in [47]. For normally distributed or constant c, the solution $X(t)$ is a Gaussian process, the so-called **Ornstein-Uhlenbeck velocity process**.

By integration of the velocity $X(t)$, we obtain the position

$$Y(t) = Y_0 + \int_0^t X(s)ds$$

of the particle. If c and Y_0 are normally distributed or constant, $Y(t)$ is, with $X(t)$, a Gaussian process, the so-called **Ornstein-Uhlenbeck (position) process**.

2.11 PDP processes

A piecewise deterministic process (PDP) is a model governed by a set of differential equations that change their deterministic structure at random points in time. PDP models define the largest class of Markov processes including most of the non-diffusion models of applied probability. This general class of deterministic-stochastic models was proposed in [36]. It has many important applications in mathematical biology, finance, and physics.

Piecewise deterministic processes appear in probability calculus and operation research, stochastic hybrid systems, reliability analysis, statistical physics, and financial mathematics; see, e.g. [2, 12, 22, 27, 31, 32, 42], for recent works and additional references.

In this thesis, we focus on processes that switch randomly between deterministic states driven by a renewal process, denoted with $\mathscr{S}(t)$, that is a discrete stochastic process. For illustration, we consider a PDP model that is a first-order system of ordinary differential equations where the known driving function of the dynamics is affected by a renewal process. The d-components state function $X(t)$, $X : [t_0, \infty) \to \Omega$, $\Omega \subseteq \mathbb{R}^d$, is defined by the following properties. We have

(a) The state function satisfies the following equation

$$\frac{d}{dt}X(t) = A_{\mathscr{S}(t)}(X), \quad t \in [t_0, \infty), \tag{2.8}$$

where $\mathscr{S}(t) : [t_0, \infty[\to \mathbb{S}$ is a Markov process (defined below in **(c)** and **(d)**) with discrete states $\mathbb{S} = \{1, \ldots, S\}$. Correspondingly, given $s \in \mathbb{S}$, we say that the dynamics is in the (deterministic) state s, and it is driven by the function $A_s : \Omega \to \mathbb{R}^d$, that belongs to the set of functions $\{A_1, \ldots, A_S\}$. We require that all $A_s(\cdot), s \in \mathbb{S}$, be Lipschitz continuous, so that for fixed s, the solution $X(t)$ exists and is unique and bounded.

(b) The state function satisfies the initial condition $X(t_0) = X_0 \in \Omega$, being in the initial state $s_0 = \mathscr{S}(t_0)$.

(c) The process $\mathscr{S}(t)$ is characterized by an exponential probability density function $\psi_s : \mathbb{R}^+ \to \mathbb{R}^+$, of transition events, as follows

$$\psi_s(t) = \mu_s e^{-\mu_s t}, \quad \text{with} \int_0^\infty \psi_s(t)\, dt = 1, \tag{2.9}$$

for each state $s \in \mathbb{S}$. In other words, it is the PDF for the time that the system stays in the state s, that is the time between consecutive events of a Poisson process.

(d) The process $\mathscr{S}(t)$ is modeled by a stochastic transition probability matrix, $\hat{q} := \{q_{ij}\}$, with the following properties

$$0 \leq q_{ij} \leq 1, \qquad \sum_{i=1}^{S} q_{ij} = 1, \qquad \forall i, j \in \mathbb{S}. \tag{2.10}$$

When a transition event occurs, the PDP system switches instantaneously from a state $j \in \mathbb{S}$, with the driving function A_j, randomly to a new state $i \in \mathbb{S}$, driven by the function A_i. Virtual transitions from the state j to itself are allowed for this model, that is, $q_{jj} > 0$.

Both **(c)** and **(d)** define the Markov renewal process $\mathscr{S}(t)$, that generates a temporal sequence of transition events $(t_0, t_1, \ldots, t_k, t_{k+1}, \ldots)$ and states $(s_0, s_1, \ldots, s_k, s_{k+1}, \ldots)$. It is said that **(d)** defines the embedded discrete Markov chain of the process $\mathscr{S}(t)$, while **(c)** defines a continuous time process for the epochs where the Markov chain changes its state, i.e. a renewal process.

The state space of the PDP process is the union of S disjoint copies of $\mathbb{R}^d$. In the j-th copy, the vector field is A_j and, at a jump time, a change to another component A_i, with i randomly chosen, can occur. Notice that the state function $X(t)$ is continuous through the events of the renewal process.

Chapter 3

Fokker-Planck equations

The origin of the name "Fokker-Planck Equation" is from the work of *Fokker* (1914) and *Planck* (1917) where the former investigated Brownian motion in a radiation field and the latter attempted to build a complete theory of fluctuations based on it. Mathematically oriented works tend to use the term "Kolmogorov's Equation" because of Kolmogorov's work in developing its rigorous basis. Yet others use the term "Smoluchowski Equation" because of Smoluchowski's original use of this equation.

In this thesis, we consider two classes of Fokker-Planck (FP) equations corresponding to Itō stochastic processes and piecewise deterministic processes. This chapter introduces these two classes of FP equations, and includes details on derivation of the equations. In the case of Itō models, we will have FP equations of parabolic type, which have been studied deeply in the literature written about stochastic models. Therefore, we present some details taken from [47]. However, regarding the FP equations corresponding to PDP processes the investigation in literature is very rare and the topic is new. In this case, we have a system of FP equations of hyperbolic type.

3.1 Itō stochastic models

3.1.1 Fokker-Planck equation in one dimension

In one dimension, the FP equation has the simple form

$$\partial_t f(x,t) = -\partial_x \left(b(x,t) f(x,t) \right) + \frac{1}{2} \partial_{xx} \left(a(x,t) f(x,t) \right),$$

for the Itō process introduced in Chapter 2, which is given by the stochastic differential equation

$$dX(t) = b(X(t),t)dt + \sigma(X(t),t)\, dW(t)$$

with drift $b(X(t),t)$, dispersion $\sigma(X(t),t) = \sqrt{a(X(t),t)}$, and Wiener process $W(t)$.

The conditional probability satisfies the FP equation, that is,

$$f(x,t) = P(x,t|x_0,t_0)$$

for any initial x_0 and t_0, and with the initial condition

$$P(x,t|x_0,t_0) = \delta(x-x_0). \tag{3.1}$$

However, using the definition for the one time probability

$$P(x,t) = \int P(x,t;x_0,t_0)\,dx_0 = \int P(x,t|x_0,t_0)P(x_0,t_0)\,dx_0,$$

we see that it is also valid for $P(x,t)$ with the initial condition

$$P(x,t)|_{t=t_0} = P(x,t_0).$$

This initial condition is generally less singular than (3.1), which is the Dirac delta function.

3.1.2 Fokker-Planck equation in several dimensions

In multidimensional cases, FP equations show more complex behavior than is possible in the case of one variable. Boundaries are no longer simple end points of a line but rather curves or surfaces, and the nature of the boundary can change from place to place [47].

If $X(t)$ is an d-dimensional random vector obeying the stochastic differential equation

$$dX(t) = b(X(t),t)\,dt + \sigma(X(t),t)\,dW(t),$$

and $W(t)$ is an m-dimensional Wiener process, then the probability density $f(x,t)$ for the random vector $X(t)$ satisfies the Fokker-Planck equation

$$\partial_t f(x,t) = -\sum_{i=1}^{d} \partial_{x_i}\left(b_i(x,t)f(x,t)\right) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \partial_{x_i x_j}\left(a_{ij}(x,t)f(x,t)\right),$$

with drift vector $b = (b_1, \ldots, b_d)$ and diffusion tensor

$$a_{ij}(x,t) = \sum_{k=1}^{m} \sigma_{ik}(x,t)\sigma_{kj}(x,t).$$

3.1.3 Boundary conditions

The FP equation is a second-order partial differential equation, and for solutions we need boundary conditions at the end points of the domain inside which x is constrained. These take on a variety forms.

We note that the FP equation

$$\partial_t f(x,t) = -\sum_{i=1}^{d} \partial_{x_i}\left(b_i(x,t)f(x,t)\right) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \partial_{x_i x_j}\left(a_{ij}(x,t)f(x,t)\right)$$

can also be written in the following form

$$\partial_t f(x,t) + \sum_{i=1}^{d} \partial_{x_i} F_i(x,t) = 0 \tag{3.2}$$

where the components of the flux F are defined as

$$F_i(x,t) = b_i(x,t) f(x,t) - \frac{1}{2} \sum_{j=1}^{d} \partial_{x_j} \left(a_{ij}(x,t) f(x,t) \right).$$

Equation (3.2) has the form of a local conservation equation, and can be written in an integral form as follows.

Consider some region R with a boundary S, and define

$$P(R,t) = \int_R f(x,t)\, dx.$$

Then (3.2) is equivalent to

$$\partial_t P(R,t) = - \int_S n \cdot F(x,t)\, dS, \tag{3.3}$$

where n is the outward pointing normal to S. Thus, (3.3) indicates that the total loss of probability of being in R is given by the surface integral of F over the boundary of R. We can now consider the different boundary conditions separately.

a) Reflecting Barrier

We can consider the situation where the process cannot leave a region R, hence there is zero net flow of probability across S, the boundary of R. Thus we require

$$n \cdot F(x,t) = 0 \quad \text{for}\, x \in S,\, n = \text{normal to}\, S.$$

Since the process cannot cross S, it must be reflected there, and hence the name *reflecting barrier* is used for this condition.

b) Absorbing Barrier

Here, one assumes that the moment the process reaches S, it is removed from the system, thus the barrier absorbs. Consequently, the probability of being on the boundary is zero, i.e.

$$f(x,t) = 0 \quad \text{for } x \in S.$$

c) Periodic Boundary Condition

We assume that the process takes place on an interval $[c,d]$ in which the two endpoints are identified with each other. This occurs, for example, if the diffusion is on a circle. Then we impose periodic boundary conditions

$$\text{(I)}: \lim_{x\to d^-} f(x,t) = \lim_{x\to c^+} f(x,t), \tag{3.4}$$

$$\text{(II)}: \lim_{x\to d^-} F(x,t) = \lim_{x\to c^+} F(x,t). \tag{3.5}$$

Most frequently, periodic boundary conditions are imposed when the functions $a(x,t)$ and $b(x,t)$ are periodic on the same interval so that we have

$$a(d,t) = a(c,t),$$

$$b(d,t) = b(c,t).$$

It means that (I) and (II) simply reduce to an equality of $f(x,t)$ and its derivatives at the points c and d.

d) Natural Boundaries

If the diffusion coefficient vanishes at a boundary, we have a situation in which the boundary may be automatically prescribed. The detailed formulation can be found in [47]. The natural boundary condition typically is applied when the range of the random variable $X(t)$ is infinite or semi-infinite. In this case $f(x,t) \to 0$ as $x \to \infty$ or $x \to -\infty$ with the decay to zero being sufficiently fast to ensure that the normalisation integral is

$$\int_{-\infty}^{\infty} f(x,t)dx = 1.$$

In the one-dimensional case, this requires that $f(x,t)$ tends to zero faster than $|x|^{-1}$ as $|x|$ tends to ∞.

e) Boundaries at infinity

All of the above kinds of boundaries can be considered at infinity, provided we can simultaneously guarantee the normalisation of the probability. If $f(x)$ is reasonably well behaved, it requires

$$\lim_{x\to\infty} f(x,t) = 0.$$

If $\partial_x f(x)$ is reasonably well behaved, that is does not oscillate infinitely rapidly as $x \to \infty$,

$$\lim_{x\to\infty} \partial_x f(x,t) = 0$$

so that a nonzero flux at infinity will usually require either $a(x,t)$ or $b(x,t)$ to become infinite there. Treatment of such cases is usually best carried out by changing to another variable which is finite at $x = \infty$.

Where there are boundaries at $x = \pm\infty$ and nonzero fluxes at infinity are permitted, we have two possibilities which do not allow for loss of probability:

i) $F(\pm\infty, t) = 0$,

ii) $F(+\infty, t) = F(-\infty, t)$.

These are the limits of reflecting and periodic boundary conditions, respectively.

3.1.4 Stationary solutions

3.1.4.1 Homogeneous Fokker-Planck equations

In a homogeneous process, the drift and diffusion coefficients are time independent. In such a case, the equation satisfied by the stationary distribution is

$$\frac{d}{dx}(b(x)f(x)) - \frac{1}{2}\frac{d^2}{dx^2}(a(x)f(x)) = 0,$$

which can also be written in terms of the flux as follows

$$\frac{d}{dx}F(x) = 0,$$

which clearly has the solution

$$F(x) = \text{constant}.$$

Suppose the process takes place on an interval (c, d). Then we must have

$$F(c) = F(x) = F(d) \equiv F, \tag{3.6}$$

and if one of the boundary conditions is reflecting, this means that both are reflecting, and $F = 0$.

If the boundaries are not reflecting, (3.6) requires them to be periodic. We then use the boundary conditions given by (3.4) and (3.5). We find the following stationary solutions corresponding to different boundary conditions.

a) Zero flux - Potential solution

Suppose that the boundary conditions are reflecting. This means zero flux trough the boundary. Setting $F = 0$, we rewrite (3.6) as follows

$$b(x)f(x) = \frac{1}{2}\frac{d}{dx}(a(x)f(x)) = 0,$$

for which the solution is

$$f(x) = \frac{\mathcal{N}}{a(x)} \exp(2\int_c^x b(s)/a(s)\,ds), \tag{3.7}$$

where $\mathcal{N}$ is a normalisation constant such that

$$\int_c^d f(x)\,dx = 1.$$

Such a solution is known as a *potential solution*, for various historical reasons, but mainly because the stationary solution is obtained by a single integration.

b) Periodic boundary condition
In this case, the flux

$$F = b(x)f(x) - \frac{1}{2}\frac{d}{dx}(a(x)f(x)). \tag{3.8}$$

is not arbitrary, but is determined by normalisation and the periodic boundary condition

$$f(c) = f(d), \tag{3.9}$$
$$F(c) = F(d). \tag{3.10}$$

For convenience, define

$$\psi(x) = \exp(2\int_c^x b(s)/a(s)\,ds).$$

Then we can easily integrate (3.8) to get

$$f(x)a(x)/\psi(x) = f(c)a(c)/\psi(c) + F\int_c^x ds/\psi(s).$$

By imposing the boundary condition (3.9), we find that

$$F = (a(d)/\psi(d) - a(c)/\psi(c))f(c)/\left(\int_c^d ds/\psi(s)\right),$$

so that

$$f(x) = f(c)\left(\frac{\displaystyle\int_c^x \frac{ds}{\psi(s)}\frac{a(d)}{\psi(d)} + \int_x^d \frac{ds}{\psi(s)}\frac{a(c)}{\psi(c)}}{\displaystyle\frac{a(x)}{\psi(x)}\int_c^d \frac{ds}{\psi(s)}}\right). \tag{3.11}$$

c) Infinite range and singular boundaries
In either of these cases, one or the other of the above possibilities may turn out to be forbidden. In general, it is very complicated to enumerate the possibilities. We shall demonstrate these by means of the following subsections.

3.1.4.2 Diffusion in a gravitational field

A strongly damped Brownian particle moving in a constant gravitational field is often described by the SDE

$$dX(t) = -b\,dt + \sqrt{a}\,dW(t),$$

for which the Fokker-Planck equation is

$$\partial_t f = \partial_x(bf) + \frac{1}{2}a\partial_{xx}f.$$

On the interval (c, d) with reflecting boundary conditions, the stationary solution is given by (3.7), that is

$$f(x) = \mathcal{N} \exp(-2bx/a),$$

where we have absorbed constant factors into the definition of $\mathcal{N}$.

Clearly this solution is normalisable on (c, d) only if c is finite, though d may be infinite. The result is no more profound than to say that particles diffusing in a beaker of fluid will fall dawn, and if the beaker is infinitely deep, they will never stop falling. Diffusion upwards against gravity is possible for any distance but with exponentially small probability.

Now assume periodic boundary conditions on (c, d). Substitution into (3.11) yields

$$f(x) = f(c);$$

a constant distribution. The interpretation is that the particles pass freely from c to d and back.

3.1.4.3 Ornstein-Uhlenbeck process

Corresponding to the Ornstein-Uhlenbeck process

$$dX(t) = -\gamma X(t)\, dt + \sqrt{a}\, dW(t)$$

we have the Fokker-Planck equation

$$\partial_t f = \partial_x(\gamma x f) + \frac{1}{2} a \partial_{xx} f,$$

whose stationary solution on the interval (c, d) with reflecting barrier is

$$f(x) = \mathcal{N} \exp(-\gamma x^2/a).$$

Provided $\gamma > 0$, this is normalisable on $(-\infty, \infty)$. If $\gamma < 0$, one can only make sense of it on a finite interval. Suppose $c = -d < 0$. If we consider periodic boundary condition on this interval, we find that

$$f(x) = f(c) \exp\left(-\frac{\gamma}{a}(x^2 - c^2)\right),$$

so that the symmetry yields the same solution as in the case of reflecting barriers.

Letting $c \to \infty$, we see that we still have the same solution. The result is also true if $c \to \infty$ independently of $d \to -\infty$, provided $\gamma > 0$.

3.1.4.4 Chemical reaction model

Although chemical reactions are normally best modelled by a birth-death master equation formalism, approximate treatments are often given by means of a Fokker-Planck equation. The reaction

$$A + B \rightleftharpoons 2A$$

is of interest since it possesses an exit boundary at $x = 0$ (where x is the number of molecules of A). Clearly if there is no A, a collision between A and B cannot occur, so no additional A is produced. The corresponding Fokker-Planck equation is [47]

$$\partial_t f = -\partial_x((cx - x^2)f(x,t)) + \frac{1}{2}\partial_{xx}((cx + x^2)f(x,t)).$$

We introduce reflecting boundaries at $x = \alpha$ and $x = \beta$. In this case, the stationary solution is

$$f(x) = \exp(-2x)(c + x)^{4\alpha-1}/x$$

which is not normalisable if $\alpha = 0$. The pole at $x = 0$ is a result of the absorption there. The stationary solution has relevance only if $\alpha > 0$, since it is otherwise not normalisable. The physical meaning of a reflecting barrier is quite simple: whenever a molecule of A disappears, we simply add another one immediately. The time for all x to disappear is in practice extraordinarily long, and the stationary solution is a good representation of the distribution except near $x = 0$.

3.1.5 Derivation of Fokker-Planck equation

In this section, we illustrate a classical heuristic construction of the Fokker-Planck equation starting from a discrete random walk; see, e.g., [33]. Let us consider the random motion of a particle that can take small steps of amount $-\Delta x$, 0 and $+\Delta x$ in an interval of time Δt. For now, we assume both Δx and Δt be fixed. Further, let us denote $\pi_{\Delta x}^{(+)}(x)$ and $\pi_{\Delta x}^{(-)}(x)$ the probabilities that the particle starting at x at time t, will be at $x + \Delta x$ and $x - \Delta x$, respectively. So that, $1 - \pi_{\Delta x}^{(+)}(x) - \pi_{\Delta x}^{(-)}(x)$ is the probability that the particle remains at x at time $t + \Delta t$. The subscript Δx of $\pi_{\Delta x}^{(+)}$ means that the probability is scale dependent and it changes as Δx approaches zero. We assume that $P(x_0, x; t)\Delta x$ is the conditional probability that the particle arrives at x at time t starting from x_0 at $t = 0$ following a random path. The following equation holds for the conditional probabilities

$$\begin{aligned} P(x_0, x; t)\Delta x &= P(x_0, x - \Delta x; t - \Delta t)\pi_{\Delta x}^{(+)}(x - \Delta x)\Delta x \\ &+ P(x_0, x + \Delta x; t - \Delta t)\pi_{\Delta x}^{(-)}(x + \Delta x)\Delta x \\ &+ P(x_0, x; t - \Delta t)(1 - \pi_{\Delta x}^{(+)}(x) - \pi_{\Delta x}^{(-)}(x))\Delta x. \end{aligned} \tag{3.12}$$

From this discrete model of a stochastic process, we want to build one with infinitesimal increments for $\Delta x, \Delta t \to 0$. In order that the limiting process has a statistical meaning, the probabilities $\pi_{\Delta x}^{(+)}$ and $\pi_{\Delta x}^{(-)}$, and the space scale Δx must be subject to some constraint while scaling $\Delta t \to 0$. These are the following infinitesimal mean of change particle position $X(t)$, conditional on $X(t) = x$,

$$b(x) = \lim_{\Delta t \to 0} \frac{E[X(t+\Delta t) - X(t)|X(t) = x]}{\Delta t} \tag{3.13}$$

and the infinitesimal variance is given by

$$a(x) = \lim_{\Delta t \to 0} \frac{V[X(t+\Delta t) - X(t)|X(t) = x]}{\Delta t}. \tag{3.14}$$

Given the particle at x at time t, then at time $t + \Delta t$ the mean value of change in position is as follows

$$\Delta x(\pi^{(+)}_{\Delta x}(x) - \pi^{(-)}_{\Delta x}(x)) \tag{3.15}$$

and the variance is given by

$$\Delta x^2(\pi^{(+)}_{\Delta x}(x) + \pi^{(-)}_{\Delta x}(x) - (\pi^{(+)}_{\Delta x}(x) - \pi^{(-)}_{\Delta x}(x))^2). \tag{3.16}$$

From (3.13), we obtain

$$b(x) = \lim_{\Delta x, \Delta t \to 0} (\pi^{(+)}_{\Delta x}(x) - \pi^{(-)}_{\Delta x}(x)) \frac{\Delta x}{\Delta t} \tag{3.17}$$

and from (3.14), we have

$$a(x) = \lim_{\Delta x, \Delta t \to 0} (\pi^{(+)}_{\Delta x}(x) - \pi^{(-)}_{\Delta x}(x) - (\pi^{(+)}_{\Delta x}(x) - \pi^{(-)}_{\Delta x}(x))^2) \frac{\Delta x^2}{\Delta t}. \tag{3.18}$$

These last two equations provide constraints for the form of $\pi^{(+)}_{\Delta x}(x)$ and $\pi^{(-)}_{\Delta x}(x)$. Here, we are building an infinitesimal stochastic process with mean $b(x)$ and variance $a(x)$. In order that $a(x)$ be a non-vanishing function and bounded $a(x) < A$, we suppose the scale law $(\Delta x)^2 = A\Delta t$. The choices

$$\pi^{(+)}_{\Delta x}(x) = \frac{1}{2A}(a(x) + b(x)\Delta x)$$

and

$$\pi^{(-)}_{\Delta x}(x) = \frac{1}{2A}(a(x) - b(x)\Delta x)$$

make the requirements on the mean and variance satisfied. Moreover, $\pi^{(+)}_{\Delta x}, \pi^{(-)}_{\Delta x} \geq 0$ and $\pi^{(+)}_{\Delta x} + \pi^{(-)}_{\Delta x} \leq 1$ must be satisfied, so that we require $a(x) \geq b(x)\Delta x$. We notice that the scaling law $\Delta x = O(\sqrt{\Delta t})$ is typical of the Wiener or Gaussian white noise.

By expanding in Taylor series (3.12) up to second order in Δt and then in Δx, we obtain

$$\begin{array}{rcl} P & \simeq & (P - P_x\Delta x + \frac{1}{2}P_{xx}\Delta x^2 - P_t\Delta t)(\pi^{(+)}_{\Delta x} - \pi^{(+)'}_{\Delta x}\Delta x + \frac{1}{2}\pi^{(+)''}_{\Delta x}\Delta x^2) \\ & + & (P + P_x\Delta x + \frac{1}{2}P_{xx}\Delta x^2 - P_t\Delta t)(\pi^{(+)}_{\Delta x} + \pi^{(+)'}_{\Delta x}\Delta x + \frac{1}{2}\pi^{(+)''}_{\Delta x}\Delta x^2) \\ & + & (P - P_t\Delta t)(1 - \pi^{(+)}_{\Delta x} - \pi^{(-)}_{\Delta x}). \end{array} \tag{3.19}$$

Finally, by using (3.17) and (3.18), and the scale law for Δx, we obtain the Fokker-Planck equation

$$\partial_t P(x_0, x; t) = \frac{1}{2}\partial_{xx}(a(x)P(x_0, x; t)) - \partial_x(b(x)P(x_0, x; t)).$$

If we insert the time in $a(x, t)$ and $b(x, t)$, there is no change in the proof and we get a more general form of the FP equation.

3.1.6 Existence and uniqueness of solutions

We introduce some assumptions on the FP model that guarantee its solvability. Let ρ be the initial PDF distribution. Consider the following Fokker-Planck model

$$\partial_t f(x,t) - \frac{1}{2}\sum_{i,j=1}^{d} \partial_{x_i x_j}\left(a_{ij}(x,t)\, f(x,t)\right) + \sum_{i=1}^{d} \partial_{x_i}\left(b_i(x,t)\, f(x,t)\right) = 0, \qquad (3.20)$$

$$f(x, t_0) = \rho(x), \qquad (3.21)$$

We have the following

Assumption 1.

1. *The coefficient function a_{ij} is bounded and satisfies the following uniform ellipticity condition for a constant $\theta > 0$,*

$$\sum_{ij=1}^{d} a_{ij}(x,t)\xi_i\xi_j \geq \theta\,|\xi|^2, \qquad \forall\, \xi \in \mathbb{R}^d,\ (x,t) \in Q.$$

2. *The coefficient functions b_i and $\partial_{x_i} a_{ij}$, $i, j = 1, \ldots, d$, satisfy the following*

$$b_i,\ \partial_{x_i} a_{ij} \in L^q(0, T; L^p(\Omega))$$

where p and q are such that $2 < p, q \leq \infty$, and $\frac{d}{2p} + \frac{1}{q} < \frac{1}{2}$.

3. *The functions $\partial_{x_i} b_i$, $i = 1, \ldots, d$, satisfy*

$$\partial_{x_i} b_i \in L^q(0, T; L^p(\Omega))$$

where p and q are such that $1 < p, q \leq \infty$, and $\frac{d}{2p} + \frac{1}{q} < 1$.

These assumptions were introduced in [11] to prove existence and uniqueness of non-negative solutions of parabolic problems. In [7], the results of [11] have been specialized to prove existence, uniqueness, and positivity of solutions to the forward FP problem (3.20)-(3.21) in a bounded domain. We have the following

Theorem 1. *Suppose that b_i and a_{ij} in (3.20) satisfy the Assumption 1, and take the initial condition $\rho \in H_0^1(\Omega)$ and homogeneous boundary conditions on $\Sigma = \partial\Omega \times (0, T)$. Then there exists a unique weak solution f to (3.20)-(3.21). Further, the solution f has the following additional property*

If $0 \leq \rho \leq m$ *a.e. in* Ω*, then*

$$0 \leq f(x,t) \leq m(1 + C\,k), \qquad \text{in } Q$$

where $k = \frac{1}{2}\sum_{i=1}^{d} \| \sum_{j=1}^{d} \partial_{x_j} a_{ij}\|_{p,q} + \| \sum_{i=1}^{d} \partial_{x_i} b_i\|_{p,q}$ *and* C *depends only on* T*,* Ω*, and the structure of the FP operator.*

For the definition of a weak solution we refer to [102]. We use Theorem 1 to discuss existence, uniqueness, and positivity of the solution to the FP problem (3.20)-(3.21) in an unbounded domain $\Omega = \mathbb{R}^d$. To this end, we define some special boundary value problems, as it is proposed in [11].

Let $\Omega^k = \{x; |x| < k\}$ and $Q^k = \Omega^k \times (0,T)$. For each integer $k \geq 3$, let $\zeta^k = \zeta^k(x)$ denote a $C_0^\infty(\mathbb{R}^n)$ function such that $\zeta^k = 1$ for $|x| \leq k-2$, $\zeta^k = 0$ for $|x| \geq k-1$, $0 \geq \zeta^k \leq 1$ and $|\partial_x \zeta^k|$ is bounded independent of k. According to Theorem 1, for each k there exists a unique and bounded weak solution f^k to the boundary value problem

$$\begin{aligned} \partial_t f^k - \tfrac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}\left(a_{ij}\, f^k\right) + \sum_{i=1}^{d} \partial_{x_i}\left(b_i(u)\, f^k\right) = 0 &\quad \text{in } Q^k, \\ f^k(x,0) = \zeta^k(x)\rho(x) &\quad \text{in } \Omega^k, \end{aligned} \tag{3.22}$$

with homogeneous boundary conditions. Extend the domain of definition of f^k by setting $f^k = 0$ for $|x| \geq k$.

In [11] one can find the arguments which prove the following theorem.

Theorem 2. *If* b_i *and* a_{ij} *in (3.20) satisfy the Assumption 1, the function* f^k *in (3.22) is bounded,* $\rho \in H^1(\Omega)$*, and* $\rho \geq 0$ *almost everywhere in* Ω*, then the problem (3.20)-(3.21) possesses a unique and non-negative weak solution*

Although the results presented in [11] and later generalized in [58] prove the existence of a unique non-negative solution for our problem belonging to the space $L^\infty((0,T); L^2_{loc}(\Omega)) \cap L^2((0,T); H^1_{loc}(\Omega))$, the arguments provided in [63] show that this solution may have higher regularity. In fact, we have that for $r > 0$, $f \in H^{r+2,r/2+1}(Q)$ as long as $\rho \in H^{r+2}(\Omega)$ and the coefficients a_{ij} and b_i belong to $H^{r,r/2}(Q)$.

Notice that in our case and in many applications the FP parameter functions are smooth and Assumption 1 and the assumptions in [63] are immediately satisfied. For additional results on the Fokker-Planck equation with irregular coefficients see [67].

3.2 Piecewise Deterministic Processes

In this thesis, we also deal with a class of PDP models described by a state function that is continuous in time and driven by a discrete state Markov process. These models are used to describe phenomena where a deterministic motion is subject to a sudden interaction that produces randomization of the motion for a short time compared to the time scale where the process is observed. In this framework, applications include dichotomic noise, random telegraph processes, transport processes, and binary noise. Further applications include reacting-diffusing systems [55], biological dispersal [54], non-Maxwellian equilibrium [5, 9, 43, 82], and filtered telegraph signal analysis [89].

We specifically consider a process in which the d-components state function $X(t)$, $X : [t_0, \infty) \to \Omega$, $\Omega \subseteq \mathbb{R}^d$, satisfies the differential equation

$$\frac{d}{dt} X(t) = A_{\mathscr{S}(t)} \left(X(t) \right), \quad t \in [t_0, \infty), \tag{3.23}$$

where $\mathscr{S}(t) : [t_0, \infty[\to \mathbb{S}$ is a Markov process with discrete states $\mathbb{S} = \{1, \dots, S\}$. For each state $s \in \mathbb{S}$, we say that the dynamics is in the state s, and it is driven by the function $A_s : \Omega \to \mathbb{R}^d$, that belongs to the set of functions $\{A_1, \dots, A_S\}$. We require that all $A_s(\cdot)$, $s \in \mathbb{S}$, be Lipschitz continuous, so that for fixed s the solution $X(t)$ exists and is unique.

In the following, we present the arguments from [8] to show that the time evolution of the probability density function of the process (3.23) is governed by the following FP hyperbolic system,

$$\partial_t f_s(x,t) + \partial_x (A_s f_s(x,t)) = \sum_{j=1}^{S} \mathcal{Q}_{sj} f_j(x,t), \qquad s \in \mathbb{S}, \tag{3.24}$$

where $\mathcal{Q}_{sj}$, $s, j \in \mathbb{S}$, are the components of the transition matrix $\mathcal{Q}$. Since the f_s, for $s \in \mathbb{S}$, represent the PDFs, we require the following

$$\sum_{s=1}^{S} \int_\Omega f_s(x,t)\, dx = 1. \tag{3.25}$$

The initial conditions for the solution of the FP system are given as follows

$$f_s(x,0) = f_s^0(x), \qquad s = 1, \dots, S, \tag{3.26}$$

where $f_s^0(x) \geq 0$, $x \in \Omega$, $\sum_{s=1}^{S} \int_\Omega f_s^0(x)\, dx = 1$.

3.2.1 Derivation of Fokker-Planck equation

We present the derivation of the one-dimensional FP equation from [8], and for multidimensional case we refer to [3].

Consider small time steps of size Δt. At each time step there is the probability $(\mu_s \Delta t)$ that a transition event occurs and $(1 - \mu_s \Delta t)$ that the jump in the process does not occur. In the latter case, the motion is deterministic. Let $\mathbb{P}$ be the probability measure on the hybrid space state of the system; then for a small interval of time Δt and displacement Δx, the change of the probability distribution $F_s(x,t) := \mathbb{P}(X(t) \leq x, \mathcal{S}(t) = s)$ for the state s is as follows

$$F_s(x + \Delta x, t + \Delta t) \simeq (1 - \mu_s \Delta t) F_s(x,t).$$

Notice that the right-hand side of this expression represents the probability that the jump in time does not occur and the evolution of the probability distribution follows a deterministic law. Next, we assume enough regularity of F_s and A_s and use Taylor expansion to obtain the following approximation

$$F_s + A_s(x)\,\Delta t\,\partial_x F_s + \partial_t F_s \Delta t \simeq (1-\mu_s \Delta t)F_s,$$

where we used the approximation $\Delta x \approx A_s(x)\Delta t$.

Now, by considering the first-order terms, we obtain the following

$$\partial_t F_s(x,t) = -A_s(x)\,\partial_x F_s(x,t) - \mu_s\,F_s(x,t). \tag{3.27}$$

This equation is valid as long as no switching event occurs. To take into account a switching event within the time interval $(t, t+\Delta t)$, we must include on the right-hand side the amount of ingoing probability from the other states. This is given by the joint probability described by the stochastic matrix q_{sj}, with $\sum_{s=1}^{S} q_{sj} = 1$, multiplied by $\mu_j\,F_j(x,t)\Delta t$. Therefore, (3.27) becomes the following

$$\partial_t F_s(x,t) + A_s(x)\,\partial_x F_s(x,t) = -\mu_s\,F_s(x,t) + \sum_{j=1}^{S} q_{sj}\mu_j F_j(x,t). \tag{3.28}$$

We refer to this equation as the Liouville master equation for the probability distribution function.

The marginal PDF f_s for the state s is related to the corresponding probability distribution as follows

$$F_s(x,t) = \int_{-\infty}^{x} f_s(z,t)\,dz, \qquad s = 1,\dots,S.$$

Hence, by differentiating (3.28) with respect to x, we obtain the following FP system for the probability density functions

$$\partial_t f_s(x,t) + \partial_x\left(A_s(x)\,f_s(x,t)\right) = \sum_{j=1}^{S} Q_{sj} f_j(x,t), \qquad s = 1,\dots,S, \tag{3.29}$$

where

$$Q_{sj} = \begin{cases} \mu_j\,q_{sj} & \text{if } j \neq s, \\ \mu_s\,(q_{ss}-1), \end{cases} \tag{3.30}$$

for $s \in \mathbb{S}$, $x \in \Omega \subset \mathbb{R}$, for the scalar process $X(t)$ in the state s. We have $\sum_{s=1}^{S} Q_{sj} = 0$.

Notice that we consider PDP processes whose PDFs represent the absolutely continuous part of the Lebesgue measure. For a rigorous derivation of the FP system for a general PDP process see [29], where an existence and uniqueness theorem is proved for the weak formulation solution of the FP system. In [29], the PDP is defined by a space-depending transition rate matrix $a(i,j,x)$ for jumps from i to j of the Markov chain, and a transition measure $d\mu(i,j,x)(dy)$ for random jumps on the deterministic dynamics. In our case, the transition rate is constant, i.e. $a(i,j,x) = \mu_i q_{ji}$, and discontinuous jumps in the values of $X(t)$ are not allowed, hence $d\mu(i,j,x)(dy) = \delta(x-y)\,dy$, where $\delta(x)$ is the δ-Dirac function.

3.2.2 Existence and uniqueness of solutions

The FP system given by (3.24) can be written in the following form

$$\partial_t f_s(x,t) + A_s(x,u_s)\, \partial_x f_s(x,t) = \sum_{j=1}^{S} \tilde{Q}_{sj}(x,u_s)\, f_j(x,t), \qquad s = 1, \dots, S, \tag{3.31}$$

where $\tilde{Q}_{sj}(x,u_s) = Q_{sj} - \delta_{sj}\partial_x A_s(x,u_s)$, with Q_{sj} depending on μ_j as defined in Eq. (3.30), and δ_{sj} denotes the Kronecker's delta. We denote with $f = (f_s)_{s=1}^S$ and $u = (u_s)_{s=1}^S$.

The model (3.31) is a first-order hyperbolic system in diagonal form, with coupling given through zero-order terms. This is a *strictly hyperbolic* model provided that the functions A_s are distinct for all $(x, u_s) \in (\Omega, U_s)$. It is well known that for first-order strictly hyperbolic PDEs, the solution to the initial value problem exisits and is unique with the same regularity as the initial data, see, e.g., [65].

Chapter 4

Finite difference discretization of FP equations

4.1 The Fokker-Planck equation for Itō processes

In this section, the Chang-Cooper discretization scheme for Fokker-Planck equations corresponding to Itō processes in bounded domains is investigated. It is shown that the Chang-Cooper scheme combined with backward first- and second-order finite differencing in time provides stable and accurate solutions that are conservative and positive. These properties are theoretically proven and validated by numerical experiments.

Consider the d-dimensional Itō stochastic process modelled by

$$\begin{cases} dX(t) = b(X(t),t)\,dt + \sigma(X(t),t)\,dW(t) \\ X(t_0) = X_0, \end{cases} \tag{4.1}$$

having an initial probability density of X_0 given by $f_0(x) \geq 0$ with $\int_\Omega f_0(x)\,dx = 1$. This stochastic process has been introduced in Chapter 2, and in Chapter 3 it was shown that the time evolution of the PDF of this stochastic process is governed by the following FP equation

$$\partial_t f(x,t) - \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}\left(a_{ij}(x,t) f(x,t)\right) + \sum_{i=1}^{d} \partial_{x_i}\left(b_i(x,t) f(x,t)\right) = 0 \tag{4.2}$$

$$f(x,0) = f_0(x) \tag{4.3}$$

where $(x,t) \in \Omega \times (0,T)$, and the diffusion coefficient is given by the positive-definite symmetric matrix $a = \sigma\,\sigma^\top$, with elements

$$a_{ij} = \sum_{k=1}^{m} \sigma_{ik}\,\sigma_{jk}.$$

We choose $\Omega \subset \mathbb{R}^d$ and $Q = \Omega \times (0,T)$; we also denote $\Sigma = \partial\Omega \times (0,T)$. Notice that in the FP model (4.2), the ‘space’ dimension corresponds to the number of components of the stochastic process.

The FP equation can be written in flux form and therefore in the case of bounded domains the condition of zero fluxes (reflection) on the boundary guarantees conservativeness of the total probability. In our analysis, we choose this type of boundary conditions. To write the FP equation (4.2) in conservative flux form, we consider the flux at (x,t) in th i-th direction as follows

$$F^i(x,t) = \frac{1}{2}\sum_{j=1}^{d} a_{ij}(x,t)\partial_{x_j} f(x,t) + \left(\frac{1}{2}\sum_{j=1}^{d} \partial_{x_j} a_{ij}(x,t) - b_i(x,t)\right) f(x,t).$$

To have a more compact notation, we define the vector B with the elements

$$B^i(x,t) = \frac{1}{2}\sum_{j=1}^{d} \partial_{x_j} a_{ij}(x,t) - b_i(x,t), \qquad 1 \le i \le d$$

and the matrix C with the elements

$$C^{ij}(x,t) = \frac{1}{2} a_{ij}(x,t).$$

Therefore the d-dimensional flux can be written as follows

$$F(x,t) = B(x,t) f(x,t) + C(x,t)\nabla f(x,t), \tag{4.4}$$

and the multi-dimensional FP equation can be expressed in the following flux form

$$\partial_t f(x,t) = \nabla \cdot F(x,t). \tag{4.5}$$

Our FP problem consists of solving (4.5) in a bounded domain with initial condition given by f_0 and zero-flux boundary conditions.

4.1.1 Discretization of the Fokker-Planck equation

We consider the FP problem in the time interval $(0,T)$. Although the discretization scheme is applicable for all bounded domains, we set $\Omega = (0,L)^d$ for the ease of numerical experiment where the FP equation corresponding to the Ornstein-Uhlenbeck process will be considered to validate the theoretical results. The time-step size is defined with $\delta t = T/M$, in which M is a positive integer, and $t^n = n\delta t, n = 0,1,\ldots,M$ are the time steps. Further, we consider a uniform mesh with mesh size $h = L/N$. We have the following spatial mesh

$$\Omega_h = \{x_{\mathrm{j}} \in \mathbb{R}^d : x_{\mathrm{j}} = \mathrm{j}\, h,\ \mathrm{j} \in \mathbb{Z}^d\} \cap \Omega,$$

where $\mathrm{j} = (j_1,\ldots,j_d)$ is a multi-index for the spatial position. The unit-coordinate vector in the ith direction is denoted with 1_i.

For grid functions u and v defined on Ω_h, we introduce the discrete L^2-scalar product

$$(u,v) = h^d \sum_{|\mathrm{j}| \le dN} u_{\mathrm{j}} v_{\mathrm{j}},$$

with associated norm $\|u\| = (u,u)^{1/2}$. We also represent $f(x_{\mathrm{j}}, t^n)$ by the approximated value f_{j}^n.

Next, we consider the following discretization scheme

$$D_t f_{\mathrm{j}}^n = \frac{1}{h}\sum_{i=1}^{d}(F_{\mathrm{j}+1_i/2}^{i,n} - F_{\mathrm{j}-1_i/2}^{i,n}), \tag{4.6}$$

where, for simplicity, we assume a diagonal diffusion matrix and therefore adopt the following simpler notation

$$C^i(x,t) = \frac{1}{2}a_{ii}(x,t), \qquad B^i(x,t) = \frac{1}{2}\partial_{x_i}a_{ii}(x,t) - b_i(x,t).$$

Within this setting, we consider first- and second-order backward time-differencing as follows

$$D_t f_{\mathrm{j}}^n = \frac{f_{\mathrm{j}}^{n+1} - f_{\mathrm{j}}^n}{\delta t}, \qquad D_t f_{\mathrm{j}}^n = \frac{3f_{\mathrm{j}}^{n+1} - 4f_{\mathrm{j}}^n + f_{\mathrm{j}}^{n-1}}{2\delta t},$$

where, for simplicity of notation, in $D_t f_{\mathrm{j}}^n$ we use the time index n to denote time differencing at t^{n+1}.

For space discretization, we need a second-order scheme which guarantees positivity of the probability density function together with conservation of the total probability. These are essential features that characterize the Chang-Cooper scheme. With this scheme, the following numerical flux in the i-th direction at the position $x_{\mathrm{j}+1_i/2}$ is constructed

$$F_{\mathrm{j}+1_i/2}^{i,n} = \left((1-\delta_{\mathrm{j}}^{i,n})B_{\mathrm{j}+1_i/2}^{i,n} + \frac{1}{h}C_{\mathrm{j}+1_i/2}^{i,n}\right) f_{\mathrm{j}+1_i}^{n+1} - \left(\frac{1}{h}C_{\mathrm{j}+1_i/2}^{i,n} - \delta_{\mathrm{j}}^{i,n}B_{\mathrm{j}+1_i/2}^{i,n}\right) f_{\mathrm{j}}^{n+1}, \tag{4.7}$$

in which

$$\delta_{\mathrm{j}}^{i,n} = \frac{1}{w_{\mathrm{j}}^{i,n}} - \frac{1}{\exp(w_{\mathrm{j}}^{i,n}) - 1}, \tag{4.8}$$

with $w_{\mathrm{j}}^{i,n} = hB_{\mathrm{j}+1_i/2}^{i,n}/C_{\mathrm{j}+1_i/2}^{i,n}$, where we assume that C^i and B^i are positive functions [23].

This formula results from the following linear convex combination of f at the points j and $\mathrm{j}+1_i$. We have

$$f_{\mathrm{j}+1_i/2}^{n+1} = (1-\delta_{\mathrm{j}}^{i,n})\, f_{\mathrm{j}+1_i}^{n+1} + \delta_{\mathrm{j}}^{i,n}\, f_{\mathrm{j}}^{n+1}, \qquad \delta_{\mathrm{j}}^{i,n} \in [0, 1/2].$$

The idea of implementing this combination was proposed by Chang and Cooper in [23] and it was motivated with the need to guarantee positive solutions that preserve equilibrium configuration. For this reason, Chang and Cooper notice that at equilibrium the numerical fluxes must be zero, $F_{\mathrm{j}+1_i/2}^{i} = 0$. Therefore, one obtains

$$\frac{f_{\mathrm{j}+1_i}^{n+1}}{f_{\mathrm{j}}^{n+1}} = \frac{\left(\frac{1}{h}C_{\mathrm{j}+1_i/2}^{i,n} - \delta_{\mathrm{j}}^{i,n}B_{\mathrm{j}+1_i/2}^{i,n}\right)}{\left[(1-\delta_{\mathrm{j}}^{i,n})B_{\mathrm{j}+1_i/2}^{i,n} + \frac{1}{h}C_{\mathrm{j}+1_i/2}^{i,n}\right]}. \tag{4.9}$$

On the other hand, if we solve $F(x_{\mathrm{j}+1_i/2}, t^{n+1}) = 0$, we have the following

$$\frac{f^{n+1}_{\mathrm{j}+1_i}}{f^{n+1}_{\mathrm{j}}} = \exp\left(- \int_{x_{\mathrm{j}}}^{x_{\mathrm{j}+1_i}} \frac{B^i(x,t^{n+1})}{C^i(x,t^{n+1})} dx_i \right) \approx \exp\left(- \frac{B^{i,n}_{\mathrm{j}+1_i/2}}{C^{i,n}_{\mathrm{j}+1_i/2}}\, h \right). \tag{4.10}$$

Comparison of (4.9) with (4.10) shows that we can choose the value of the parameter $\delta^{i,n}_{\mathrm{j}}$ such that (4.9) gives the exact ratio value (4.10). Thus, we obtain (4.8) where $\delta^{i,n}_{\mathrm{j}}$ can be shown to be monotonically decreasing from 1/2 to 0 as $w^{i,n}_{\mathrm{j}}$ goes from 0 to ∞; see [23]. Notice that with the choice of $\delta^{i,n}_{\mathrm{j}}$ given above, the resulting scheme shares the same properties of the continuous FP equation that guarantee positiveness and conservativeness.

4.1.1.1 The Chang-Cooper scheme with first-order time differencing

In our discussion, we first focus on the numerical solution of the FP equation in one spatial dimension and later generalize the results for $d > 1$. The one-dimensional FP problem reads as follows

$$\begin{array}{rcll} \partial_t f(x,t) &=& \partial_x F(x,t) + g(x,t), & (x,t) \in Q, \\ F(x,t) &=& 0, & (x,t) \in \Sigma, \\ f(x,0) &=& f_0(x), & x \in \Omega. \end{array} \tag{4.11}$$

where $F(x,t) = [B(x,t)f(x,t) + C(x,t)\partial_x f(x,t)]$ represents the flux. We assume that C is a positive continuous scalar function, and B satisfies Lipschitz continuity, that is, $|B(x+h,t) - B(x,t)| \leq \gamma h$, where $\gamma > 0$ is the Lipschitz constant. In particular, notice that in the case of an Ornstein-Uhlenbeck process, $C(x,t)$ is a positive constant function and $B(x,t)$ is constant in time and linear in the space variable. Therefore in this case the global growth conditions and the Lipschitz condition given in [18] are satisfied.

For convenience of the numerical analysis that follows, we consider a source term g added to the equation. However, positivity and conservativeness of the FP equation are claimed for $g = 0$.

In their work [23], Chang and Cooper proposed the following discretization of the one-dimensional FP equation with first-order implicit time-differencing

$$\begin{array}{rcl} \dfrac{f^{n+1}_j - f^n_j}{\delta t} &=& \dfrac{1}{h}\Big\{ \left[(1-\delta^n_j) B^n_{j+\frac{1}{2}} + \dfrac{1}{h} C^n_{j+\frac{1}{2}} \right] f^{n+1}_{j+1} \\ &-& \left[\dfrac{1}{h}\left(C^n_{j+\frac{1}{2}} + C^n_{j-\frac{1}{2}} \right) + (1-\delta^n_{j-1}) B^n_{j-\frac{1}{2}} - \delta^n_j B^n_{j+\frac{1}{2}} \right] f^{n+1}_j \\ &+& \left[\dfrac{1}{h} C^n_{j-\frac{1}{2}} - \delta^n_{j-1} B^n_{j-\frac{1}{2}} \right] f^{n+1}_{j-1} \Big\} \\ &+& g^{n+1}_j, \qquad\qquad j = 0, \ldots, N, \end{array} \tag{4.12}$$

with the following zero-flux boundary conditions

$$F^n_{-\frac{1}{2}} = 0, \qquad F^n_{N+\frac{1}{2}} = 0, \tag{4.13}$$

where

$$F^n_{j+\frac{1}{2}} = B^n_{j+\frac{1}{2}}\left((1-\delta^n_j)f^{n+1}_{j+1} + \delta^n_j f^{n+1}_j\right) + C^n_{j+\frac{1}{2}}\left(\frac{f^{n+1}_{j+1} - f^{n+1}_j}{h}\right).$$

Notice that in order to define the numerical zero-flux boundary conditions, ghost points and ghost variables at $j=-1$ and $j=N+1$ must be formally introduced. However, we show later how these variables are not required.

In the sequel, we investigate the approximation properties of this scheme, that we call the CC-BDF scheme. Conservativeness of the FP equation derives straightforwardly from the discrete FP equation in flux form. We have

Lemma 1. *The following conservativeness property holds*

$$\sum_{j=0}^{N} f^{n+1}_j = \sum_{j=0}^{N} f^n_j, \qquad n \geq 0.$$

Proof. The claim can be proved by summing over j in the equation (4.6) with $D_t f^n_{\mathrm{j}} = \dfrac{f^{n+1}_{\mathrm{j}} - f^n_{\mathrm{j}}}{\delta t}$, which leads to

$$\sum_{j=0}^{N}(f^{n+1}_{\mathrm{j}} - f^n_{\mathrm{j}}) = \frac{\delta t}{h}\sum_{j=0}^{N}(F^n_{j+\frac{1}{2}} - F^n_{j-\frac{1}{2}}).$$

The right-hand side vanishes since this is the difference of the fluxes at the boundaries, therefore $\sum_{j=0}^{N} f^{n+1}_{\mathrm{j}} = \sum_{j=0}^{N} f^n_{\mathrm{j}}$. □

In order to investigate stability of the CC scheme with first-order time differencing, we define the following

$$D_+ f^n_j = (f^n_{j+1} - f^n_j)/h,$$
$$D_- f^n_j = (f^n_j - f^n_{j-1})/h,$$
$$M_\delta f^n_j = (1-\delta^{n-1}_{j-1})f^n_j + \delta^{n-1}_{j-1} f^n_{j-1}.$$

With this setting, the CC-BDF scheme becomes

$$\frac{f^{n+1}_j - f^n_j}{\delta t} = D_+ C^n_{j-\frac{1}{2}} D_- f^{n+1}_j + D_+ B^n_{j-\frac{1}{2}} M_\delta f^{n+1}_j + g^{n+1}_j, \tag{4.14}$$

where

$$\begin{aligned}
D_+ C^n_{j-\frac{1}{2}} D_- f^{n+1}_j &= D_+ C^n_{j-\frac{1}{2}}\left(\frac{f^{n+1}_j - f^{n+1}_{j-1}}{h}\right)\\
&= \frac{1}{h}\left\{C^n_{j+\frac{1}{2}}\left(\frac{f^{n+1}_{j+1} - f^{n+1}_j}{h}\right) - C^n_{j-\frac{1}{2}}\left(\frac{f^{n+1}_j - f^{n+1}_{j-1}}{h}\right)\right\}\\
&= \frac{1}{h}\left\{\frac{1}{h}C^n_{j+\frac{1}{2}} f^{n+1}_{j+1} - \frac{1}{h}\left(C^n_{j+\frac{1}{2}} + C^n_{j-\frac{1}{2}}\right) f^{n+1}_j + \frac{1}{h}C^n_{j-\frac{1}{2}} f^{n+1}_{j-1}\right\},
\end{aligned}$$

and

$$
\begin{aligned}
D_+ B^n_{j-\frac{1}{2}} M_\delta f^{n+1}_j &= D_+ \left((1-\delta^n_{j-1}) B^n_{j-\frac{1}{2}} f^{n+1}_j + \delta^n_{j-1} B^n_{j-\frac{1}{2}} f^{n+1}_{j-1} \right) \\
&= \frac{1}{h} \left\{ (1-\delta^n_j) B^n_{j+\frac{1}{2}} f^{n+1}_{j+1} - (1-\delta^n_{j-1}) B^n_{j-\frac{1}{2}} f^{n+1}_j \right\} \\
&+ \frac{1}{h} \left\{ \delta^n_j B^n_{j+\frac{1}{2}} f^{n+1}_j - \delta^n_{j-1} B^n_{j-\frac{1}{2}} f^{n+1}_{j-1} \right\}.
\end{aligned}
$$

Theorem 3. *If $\delta t \le \frac{1}{2\gamma}$, then we have the following bound for the solution of the discretization scheme (4.14).*

$$
\|f^k\| \le 2^{k/2} \|f^0\| + \delta t \sum_{n=0}^{k-1} 2^{\frac{k-n+1}{2}} \|g^{n+1}\|, \quad k = 1, \cdots, M.
$$

Proof. Taking discrete L^2 inner product of (4.14) with f^{n+1}, we have

$$
\left(\frac{f^{n+1} - f^n}{\delta t}, f^{n+1} \right) = \left(D_+ C^n_{-\frac{1}{2}} D_- f^{n+1}, f^{n+1} \right) + \left(D_+ B^n_{-\frac{1}{2}} M_\delta f^{n+1}, f^{n+1} \right) + \left(g^{n+1}, f^{n+1} \right).
$$

Next, we find upper bounds for the terms on the right-hand side.

$$
\begin{aligned}
\left(D_+ C^n_{-\frac{1}{2}} D_- f^{n+1}, f^{n+1} \right) &= \textstyle\sum_{j=0}^N (D_+ C^n_{j-\frac{1}{2}} D_- f^{n+1}_j) f^{n+1}_j h \\
&= \textstyle\sum_{j=0}^N \left[C^n_{j+\frac{1}{2}} \left(\frac{f^{n+1}_{j+1} - f^{n+1}_j}{h} \right) f^{n+1}_j - C^n_{j-\frac{1}{2}} \left(\frac{f^{n+1}_j - f^{n+1}_{j-1}}{h} \right) f^{n+1}_j \right] \\
&= \textstyle\sum_{j=1}^{N+1} C^n_{j-\frac{1}{2}} \left(\frac{f^{n+1}_j - f^{n+1}_{j-1}}{h} \right) f^{n+1}_{j-1} \\
&\quad - \textstyle\sum_{j=0}^N C^n_{j-\frac{1}{2}} \left(\frac{f^{n+1}_j - f^{n+1}_{j-1}}{h} \right) f^{n+1}_j \\
&= - \textstyle\sum_{j=1}^N C^n_{j-\frac{1}{2}} \left(\frac{f^{n+1}_j - f^{n+1}_{j-1}}{h} \right) \left(\frac{f^{n+1}_j - f^{n+1}_{j-1}}{h} \right) h \\
&\quad + C^n_{N+\frac{1}{2}} \left(\frac{f^{n+1}_{N+1} - f^{n+1}_N}{h} \right) f^{n+1}_N - C^n_{-\frac{1}{2}} \left(\frac{f^{n+1}_0 - f^{n+1}_{-1}}{h} \right) f^{n+1}_0 \\
&\le C^n_{N+\frac{1}{2}} \left(\frac{f^{n+1}_{N+1} - f^{n+1}_N}{h} \right) f^{n+1}_N - C^n_{-\frac{1}{2}} \left(\frac{f^{n+1}_0 - f^{n+1}_{-1}}{h} \right) f^{n+1}_0
\end{aligned}
$$

Further, we have

$$\begin{aligned}
\left(D_+B^n_{-\frac{1}{2}}M_\delta f^{n+1}, f^{n+1}\right) &= \textstyle\sum_{j=0}^{N}((1-\delta_j^n)B^n_{j+\frac{1}{2}}f^{n+1}_{j+1}f^{n+1}_j - (1-\delta^n_{j-1})B^n_{j-\frac{1}{2}}(f^{n+1}_j)^2 \\
&\quad +\delta^n_j B^n_{j+\frac{1}{2}}(f^{n+1}_j)^2 - \delta^n_{j-1}B^n_{j-\frac{1}{2}}f^{n+1}_{j-1}f^{n+1}_j) \\
&= \textstyle\sum_{j=0}^{N}(1-\delta_j^n)B^n_{j+\frac{1}{2}}f^{n+1}_{j+1}f^{n+1}_j - \sum_{j=-1}^{N-1}\delta^n_j B^n_{j+\frac{1}{2}}f^{n+1}_j f^{n+1}_{j+1} \\
&\quad +\textstyle\sum_{j=0}^{N}\delta^n_j B^n_{j+\frac{1}{2}}(f^{n+1}_j)^2 + \sum_{j=-1}^{N-1}(\delta^n_j - 1)B^n_{j+\frac{1}{2}}(f^{n+1}_{j+1})^2 \\
&= \textstyle\sum_{j=0}^{N-1}(1-\delta_j^n)B^n_{j+\frac{1}{2}}f^{n+1}_{j+1}f^{n+1}_j - \sum_{j=0}^{N-1}\delta^n_j B^n_{j+\frac{1}{2}}f^{n+1}_j f^{n+1}_{j+1} \\
&\quad +(1-\delta^n_N)B^n_{N+\frac{1}{2}}f^{n+1}_{N+1}f^{n+1}_N - \delta^n_{-1}B^n_{-\frac{1}{2}}f^{n+1}_{-1}f^{n+1}_0 \\
&\quad +\textstyle\sum_{j=0}^{N-1}\delta^n_j B^n_{j+\frac{1}{2}}(f^{n+1}_j)^2 + \sum_{j=0}^{N-1}(\delta^n_j - 1)B^n_{j+\frac{1}{2}}(f^{n+1}_{j+1})^2 \\
&\quad +\delta^n_N B^n_{N+\frac{1}{2}}(f^{n+1}_N)^2 + (\delta^n_{-1}-1)B^n_{-\frac{1}{2}}(f^{n+1}_0)^2.
\end{aligned}$$

Now, we employ the zero-flux boundary conditions $F^n_{-\frac{1}{2}} = 0$ and $F^n_{N+\frac{1}{2}} = 0$ given by

$$B^n_{-\frac{1}{2}}\left((1-\delta^n_{-1})f^{n+1}_0 + \delta^n_{-1}f^{n+1}_{-1}\right) + C^n_{-\frac{1}{2}}\left(\frac{f^{n+1}_0 - f^{n+1}_{-1}}{h}\right) = 0,$$

and

$$B^n_{N+\frac{1}{2}}\left((1-\delta^n_{N})f^{n+1}_{N+1} + \delta^n_{N}f^{n+1}_{N}\right) + C^n_{N+\frac{1}{2}}\left(\frac{f^{n+1}_{N+1} - f^{n+1}_{N}}{h}\right) = 0.$$

We obtain

$$\begin{aligned}
&\left(D_+C^n_{-\frac{1}{2}}D_-f^{n+1}, f^{n+1}\right) + \left(D_+B^n_{-\frac{1}{2}}M_\delta f^{n+1}, f^{n+1}\right) \\
\leq\ & \textstyle\sum_{j=0}^{N-1}\left[(f^{n+1}_j)^2 + (f^{n+1}_{j+1})^2\right]B^n_{j+\frac{1}{2}}(\frac{1-2\delta^n_j}{2}) \\
&+\textstyle\sum_{j=0}^{N-1}\delta^n_j B^n_{j+\frac{1}{2}}(f^{n+1}_j)^2 + \sum_{j=0}^{N-1}(\delta^n_j - 1)B^n_{j+\frac{1}{2}}(f^{n+1}_{j+1})^2 \\
\leq\ & \textstyle\sum_{j=0}^{N-1}\frac{1}{2}B^n_{j+\frac{1}{2}}(f^{n+1}_j)^2 - \sum_{j=0}^{N-1}\frac{1}{2}B^n_{j+\frac{1}{2}}(f^{n+1}_{j+1})^2 \\
\leq\ & \textstyle\sum_{j=0}^{N-1}\frac{1}{2}B^n_{j+\frac{1}{2}}(f^{n+1}_j)^2 - \sum_{j=1}^{N}\frac{1}{2}B^n_{j-\frac{1}{2}}(f^{n+1}_j)^2 \\
&+\tfrac{1}{2}B^n_{N+\frac{1}{2}}(f^{n+1}_N)^2 - \tfrac{1}{2}B^n_{-\frac{1}{2}}(f^{n+1}_0)^2 \\
=\ & \textstyle\sum_{j=0}^{N}\frac{1}{2}B^n_{j+\frac{1}{2}}(f^{n+1}_j)^2 - \sum_{j=0}^{N}\frac{1}{2}B^n_{j-\frac{1}{2}}(f^{n+1}_j)^2 \\
\leq\ & \textstyle\sum_{j=0}^{N}\frac{1}{2}|B^n_{j+\frac{1}{2}} - B^n_{j-\frac{1}{2}}||f^{n+1}_j|^2 \\
\leq\ & \tfrac{1}{2}\gamma\textstyle\sum_{j=0}^{N}|f^{n+1}_j|^2 h \\
=\ & \tfrac{1}{2}\gamma\|f^{n+1}\|^2.
\end{aligned}$$

Therefore, we have the following estimate

$$\left(\frac{f^{n+1}-f^n}{\delta t}, f^{n+1}\right) \leq \frac{1}{2}\gamma\|f^{n+1}\|^2 + \|g^{n+1}\|\|f^{n+1}\|.$$

On the other hand, we have

$$\frac{1}{2\delta t}(\|f^{n+1}\|^2 - \|f^n\|^2) \leq \left(\frac{f^{n+1}-f^n}{\delta t}, f^{n+1}\right),$$

which is easily proved by considering $f^{n+1} = \frac{1}{2}(f^{n+1}-f^n) + \frac{1}{2}(f^{n+1}+f^n)$. Hence, we obtain

$$\frac{1}{2\delta t}(\|f^{n+1}\|^2 - \|f^n\|^2) \leq \frac{1}{2}\gamma\|f^{n+1}\|^2 + \|g^{n+1}\|\|f^{n+1}\|.$$

Since $\delta t \leq \frac{1}{2\gamma}$, we have $\frac{1}{1-\gamma\delta t} \leq 2$, and consequently

$$\|f^{n+1}\|^2 \leq 2\|f^n\|^2 + 4\delta t\|g^{n+1}\|\|f^{n+1}\|,$$

$$\|f^{n+1}\|^2 - 4\delta t\|g^{n+1}\|\|f^{n+1}\| + 4\delta t^2\|g^{n+1}\|^2 \leq 2\|f^n\|^2 + 4\delta t^2\|g^{n+1}\|^2,$$

$$\left(\|f^{n+1}\| - 2\delta t\|g^{n+1}\|\right)^2 \leq \left(\sqrt{2}\|f^n\| + 2\delta t\|g^{n+1}\|\right)^2,$$

$$\|f^{n+1}\| \leq \sqrt{2}\|f^n\| + 4\delta t\|g^{n+1}\|.$$

This recursion relation gives us the following bound

$$\|f^k\| \leq 2^{k/2}\|f^0\| + \delta t\sum_{n=0}^{k-1} 2^{\frac{k-n+1}{2}}\|g^{n+1}\|, \quad k = 1, \cdots, M.$$

□

To investigate the order of accuracy of the CC-BDF scheme, we assume $f \in C^3([0,T], C^4(\Omega))$ and estimate the size of the truncation error.

Lemma 2. *The truncation error of the discretization scheme (4.14) is of order $O(h^2 + \delta t)$.*

Proof. First we notice that

$$\frac{f_j^{n+1}-f_j^n}{\delta t} = \partial_t f_j^{n+1} - \frac{\delta t}{2}\frac{\partial^2}{\partial t^2}f_j^{n+1} + \frac{\delta t^2}{6}\frac{\partial^3}{\partial t^3}f_j^{n+1} + O(\delta t^3).$$

Therefore, we have

$$\partial_t f_j^{n+1} - \frac{f_j^{n+1}-f_j^n}{\delta t} = \frac{\delta t}{2}\frac{\partial^2}{\partial t^2}f_j^{n+1} + O(\delta t^2),$$

and also

$$\begin{aligned}
D_+C^n_{j-\frac12}D_-f^{n+1}_j &= \frac{1}{h}\left\{C^n_{j+\frac12}\left(\frac{f^{n+1}_{j+1}-f^{n+1}_j}{h}\right)-C^n_{j-\frac12}\left(\frac{f^{n+1}_j-f^{n+1}_{j-1}}{h}\right)\right\}\\
&= \frac{1}{h}\partial_x f^{n+1}_j(C^n_{j+\frac12}-C^n_{j-\frac12})+\tfrac12\partial_{xx}f^{n+1}_j(C^n_{j+\frac12}+C^n_{j-\frac12})\\
&+ \frac{h}{6}\tfrac{\partial^3}{\partial x^3}f^{n+1}_j(C^n_{j+\frac12}-C^n_{j-\frac12})+\frac{h^2}{24}\tfrac{\partial^4}{\partial x^4}f^{n+1}_j(C^n_{j+\frac12}+C^n_{j-\frac12})+O(h^3)\\
&= \partial_x f^{n+1}_j\partial_x C^n_j+\frac{h^2}{24}\partial_x f^{n+1}_j\tfrac{\partial^3}{\partial x^3}C^n_j+\tfrac{\partial^2}{\partial x^2}f^{n+1}_j C^n_j\\
&+ \frac{h^2}{8}\tfrac{\partial^2}{\partial x^2}f^{n+1}_j\tfrac{\partial^2}{\partial x^2}C^n_j+\frac{h^2}{6}\tfrac{\partial^3}{\partial x^3}f^{n+1}_j\partial_x C^n_j+\frac{h^2}{12}C^n_j\tfrac{\partial^4}{\partial x^4}f^{n+1}_j+O(x^4).
\end{aligned}$$

Therefore, we obtain the following intermediate estimate

$$\begin{aligned}
&D_+C^n_{j-\frac12}D_-f^{n+1}_j-\partial_x(C^n\partial_x f)^{t_{n+1}}_{x_j}\\
&=\frac{h^2}{2}\left(\tfrac{1}{12}\partial_x f^{n+1}_j\tfrac{\partial^3}{\partial x^3}C^n_j+\tfrac14\tfrac{\partial^2}{\partial x^2}f^{n+1}_j\tfrac{\partial^2}{\partial x^2}C^n_j+\tfrac16\tfrac{\partial^3}{\partial x^3}f^{n+1}_j\partial_x C^n_j+\tfrac16 C^n_j\tfrac{\partial^4}{\partial x^4}f^{n+1}_j\right)\\
&+O(h^4).
\end{aligned}$$

We also have

$$\begin{aligned}
&D_+B^n_{j-\frac12}M_\delta f^{n+1}_j\\
=\;&\frac{1}{h}\left\{(1-\delta^n_j)B^n_{j+\frac12}f^{n+1}_{j+1}-(1-\delta^n_{j-1})B^n_{j-\frac12}f^{n+1}_j\right\}\\
+\;&\frac{1}{h}\left\{\delta^n_jB^n_{j+\frac12}f^{n+1}_j-\delta^n_{j-1}B^n_{j-\frac12}f^{n+1}_{j-1}\right\}\\
=\;&\frac{1}{h}\left\{B^n_{j+\frac12}\left((1-\delta^n_j)f^{n+1}_{j+1}+\delta^n_jf^{n+1}_j\right)-B^n_{j-\frac12}\left((1-\delta^n_{j-1})f^{n+1}_j+\delta^n_{j-1}f^{n+1}_{j-1}\right)\right\}\\
=\;&\frac{1}{h}f^{n+1}_j(B^n_{j+\frac12}-B^n_{j-\frac12})+\partial_x f^{n+1}_j\left((1-\delta^n_j)B^n_{j+\frac12}+\delta^n_{j-1}B^n_{j-\frac12}\right)\\
+\;&\frac{h}{2}\partial_{xx}f^{n+1}_j\left((1-\delta^n_j)B^n_{j+\frac12}-\delta^n_{j-1}B^n_{j-\frac12}\right)+O(h^2)\\
=\;&f^{n+1}_j\partial_x B^n_j+\partial_x f^{n+1}_jB^n_j(1-\delta^n_j+\delta^n_{j-1})+\tfrac{h}{2}\partial_x(f^{n+1}_jB^n_j)(1-\delta^n_j-\delta^n_{j-1})\\
+\;&\tfrac{h}{2}\partial_{xx}f^{n+1}_jB^n_j(1-\delta^n_j-\delta^n_{j-1})+O(h^2).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
D_+B^n_{j-\frac12}M_\delta f^{n+1}_j-\partial_x(B^n\,f)^{t_{n+1}}_{x_j} &= \partial_x f^{n+1}_jB^n_j(\delta^n_{j-1}-\delta^n_j)\\
&+\tfrac{h}{2}\partial_x(\partial_x f^{n+1}_jB^n_j)(1-\delta^n_j-\delta^n_{j-1})+O(h^2).
\end{aligned}$$

Now, we note that

$$\delta^n_j=\frac{1}{w_j}-\frac{1}{\exp w_j-1},$$

where $w_j = h\frac{B^n_{j+\frac{1}{2}}}{C^n_{j+\frac{1}{2}}}$.

Since

$$\delta^n_j = \frac{\sum_{k=1}^{\infty}\frac{w_j^k}{(k+1)!}}{\sum_{k=1}^{\infty}\frac{w_j^k}{k!}},$$

we have

$$\begin{aligned}
\delta^n_{j-1} - \delta^n_j &= \frac{\sum_{k=1}^{\infty}\frac{w_{j-1}^k}{(k+1)!}}{\sum_{k=1}^{\infty}\frac{w_{j-1}^k}{k!}} - \frac{\sum_{k=1}^{\infty}\frac{w_j^k}{(k+1)!}}{\sum_{k=1}^{\infty}\frac{w_j^k}{k!}} \\
&= \frac{\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_{j-1}^p w_j^q}{(p+1)!q!} - \sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_{j-1}^p w_j^q}{p!(q+1)!}}{\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_{j-1}^p w_j^q}{p!q!}} \\
&= \frac{\frac{1}{12}w_{j-1}w_j(w_{j-1} - w_j) + O(h^4)}{\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_{j-1}^p w_j^q}{p!q!}},
\end{aligned}$$

and

$$\begin{aligned}
1 - \delta^n_j - \delta^n_{j-1} &= 1 - \frac{\sum_{k=1}^{\infty}\frac{w_j^k}{(k+1)!}}{\sum_{k=1}^{\infty}\frac{w_j^k}{k!}} - \frac{\sum_{k=1}^{\infty}\frac{w_{j-1}^k}{(k+1)!}}{\sum_{k=1}^{\infty}\frac{w_{j-1}^k}{k!}} \\
&= \frac{\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_j^p w_{j-1}^q}{p!q!} - \sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_j^p w_{j-1}^q}{(p+1)!q!} - \sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_{j-1}^p w_j^q}{(p+1)!q!}}{\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_{j-1}^p w_j^q}{p!q!}} \\
&= -\frac{\frac{1}{12}w_{j-1}w_j(w_{j-1} + w_j) + O(h^4)}{\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\frac{w_{j-1}^p w_j^q}{p!q!}},
\end{aligned}$$

Notice that

$$w_{j-1} - w_j = h(\frac{B^n_{j-\frac{1}{2}}}{C^n_{j-\frac{1}{2}}} - \frac{B^n_{j+\frac{1}{2}}}{C^n_{j+\frac{1}{2}}}) = O(h^2),$$

and

$$w_{j-1} + w_j = h(\frac{B^n_{j-\frac{1}{2}}}{C^n_{j-\frac{1}{2}}} + \frac{B^n_{j+\frac{1}{2}}}{C^n_{j+\frac{1}{2}}}) = O(h),$$

we obtain

$$\delta^n_{j-1} - \delta^n_j = O(h^2),$$

and

$$1 - \delta^n_j - \delta^n_{j-1} = O(h).$$

Using (4.11), the truncation error can be written as follows

$$\begin{aligned}
\phi^{n+1}_j &= [\partial_t f^{n+1}_j - \frac{f^{n+1}_j - f^n_j}{\delta t}] + [D_+ C^n_{j-\frac{1}{2}} D_- f^{n+1}_j - \partial_x(C^n \partial_x f)^{t_{n+1}}_{x_j}] \\
&+ [D_+ B^n_{j-\frac{1}{2}} M_\delta f^{n+1}_j - \partial_x(B^n f)^{t_{n+1}}_{x_j}].
\end{aligned}$$

Hence, the truncation error is $O(h^2 + \delta t)$. □

Next, we define the global error as follows

$$e_j^n = f(x_j, t^n) - f_j^n, \qquad j = 0, \dots, N,$$

and investigate the accuracy of the method by the following theorem.

Theorem 4. *If $\delta t \leq \frac{1}{2\gamma}$, then the discretization scheme (4.14) converges with an error of order $O(h^2 + \delta t)$.*

Proof. By definition of the truncation error, we have

$$\frac{e_j^{n+1} - e_j^n}{\delta t} = D_+ C_{j-\frac{1}{2}}^n D_- e_j^{n+1} + D_+ B_{j-\frac{1}{2}}^n M_\delta e_j^{n+1} + \phi_j^{n+1}.$$

That is, the solution error satisfies the discretized FP equation discussed above with the right-hand side given by the truncation error function. Hence, Theorem 3 implies accuracy as follows

$$\|e^k\| \leq \delta t \sum_{n=0}^{k-1} 2^{\frac{k-n+1}{2}} \|\phi^{n+1}\|.$$

Therefore the difference scheme (4.14) converges with error $O(h^2 + \delta t)$. □

We now notice that the fundamental theorem of numerical analysis, also known as the Lax-Richtmyer theorem [66], which states that for consistent numerical approximations stability and convergence are equivalent, proves the stability of the discretization scheme (4.14).

In order to investigate the positivity property of the CC-BDF scheme, we write the discrete FP equation in the following form

$$-\tilde{A}_j f_{j+1}^{n+1} + \tilde{B}_j f_j^{n+1} - \tilde{C}_j f_{j-1}^{n+1} = f_j^n, \qquad j = 0, ..., N, \tag{4.15}$$

where

$$\begin{aligned} \tilde{A}_j &= \tfrac{\delta t}{h^2} C_{j+\frac{1}{2}}^n W_j \exp w_j, \\ \tilde{B}_j &= \tfrac{\delta t}{h^2} (C_{j+\frac{1}{2}}^n W_j + C_{j-\frac{1}{2}}^n W_{j-1} \exp w_{j-1}) + 1, \\ \tilde{C}_j &= \tfrac{\delta t}{h^2} C_{j-\frac{1}{2}}^n W_{j-1}, \end{aligned} \tag{4.16}$$

and

$$W_j = w_j / (\exp w_j - 1). \tag{4.17}$$

By summing (4.15) over j, we have

$$-\sum_{j=0}^N \tilde{A}_j f_{j+1}^{n+1} + \sum_{j=0}^N \tilde{B}_j f_j^{n+1} - \sum_{j=0}^N \tilde{C}_j f_{j-1}^{n+1} = \sum_{j=0}^N f_j^n.$$

The left-hand side of the equation above can be simplified as follows

$$
\begin{aligned}
&-\textstyle\sum_{j=0}^{N} \tilde{A}_j f_{j+1}^{n+1} + \sum_{j=0}^{N} \tilde{B}_j f_j^{n+1} - \sum_{j=0}^{N} \tilde{C}_j f_{j-1}^{n+1} \\
&= -\textstyle\sum_{j=1}^{N+1} \tilde{A}_{j-1} f_j^{n+1} + \sum_{j=0}^{N} \tilde{B}_j f_j^{n+1} - \sum_{j=-1}^{N-1} \tilde{C}_{j+1} f_j^{n+1} \\
&= \textstyle\sum_{j=1}^{N-1} (-\tilde{A}_{j-1} + \tilde{B}_j - \tilde{C}_{j+1}) f_j^{n+1} - \tilde{A}_{N-1} f_N^{n+1} - \tilde{A}_N f_{N+1}^{n+1} \\
&+ \tilde{B}_0 f_0^{n+1} + \tilde{B}_N f_N^{n+1} - \tilde{C}_0 f_{-1}^{n+1} - \tilde{C}_1 f_0^{n+1} \\
&= \textstyle\sum_{j=1}^{N-1} \left(1 + \frac{\delta t}{h^2}(C_{j+\frac{1}{2}} W_j + C_{j-\frac{1}{2}} W_{j-1} \exp w_{j-1} - C_{j-\frac{1}{2}} W_{j-1} \exp w_{j-1} - C_{j+\frac{1}{2}} W_j)\right) f_j^{n+1} \\
&+ (\tilde{B}_N - \tilde{A}_{N-1}) f_N^{n+1} + (\tilde{B}_0 - \tilde{C}_1) f_0^{n+1} - \tilde{A}_N f_{N+1}^{n+1} - \tilde{C}_0 f_{-1}^{n+1} \\
&= \textstyle\sum_{j=0}^{N} f_j^{n+1} + \tilde{A}_N \left(f_N^{n+1} \exp(-w_N) - f_{N+1}^{n+1}\right) + \tilde{C}_0 \left(f_0^{n+1} \exp w_{-1} - f_{-1}^{n+1}\right).
\end{aligned}
$$

The conservation property is given, if the following holds

$$\tilde{A}_N \left(f_N^{n+1} \exp(-w_N) - f_{N+1}^{n+1}\right) = 0,$$

and

$$\tilde{C}_0 \left(f_0^{n+1} \exp w_{-1} - f_{-1}^{n+1}\right) = 0.$$

On the other hand, the boundary fluxes can be written as follows

$$F_{-\frac{1}{2}}^n = \frac{h}{\delta t} \tilde{C}_0 \left(f_0^{n+1} \exp w_{-1} - f_{-1}^{n+1}\right),$$

and

$$F_{N+\frac{1}{2}}^n = -\frac{h}{\delta t} \tilde{A}_N \left(f_N^{n+1} \exp(-w_N) - f_{N+1}^{n+1}\right).$$

Therefore, the zero-flux boundary conditions read as follows

$$f_{-1}^{n+1} = f_0^{n+1} \exp w_{-1}, \quad \text{and} \quad f_{N+1}^{n+1} = f_N^{n+1} \exp(-w_N),$$

which is consistent with the conservativeness of the discretization scheme.

We notice that the resulting tridiagonal problem can be written in the following matrix form

$$\mathcal{A}\, f^{n+1} = f^n,$$

with the matrix of coefficients given by the following

$$
\mathcal{A}_{ij} = \begin{cases}
-\tilde{A}_i, & j = i+1, \quad 0 \le i \le N-1, \\
\breve{B}_i, & j = i, \qquad\quad\; 0 \le i \le N, \\
-\tilde{C}_i, & j = i-1, \qquad 1 \le i \le N, \\
0, & \text{otherwise},
\end{cases}
\tag{4.18}
$$

where $\breve{B}_i = \tilde{B}_i$ for $1 \le i \le N-1$, and

$$\breve{B}_0 = \tilde{B}_0 - \tilde{C}_0 \, \exp(w_{-1}) \text{ and } \breve{B}_N = \tilde{B}_N - \tilde{A}_N \, \exp(-w_N),$$

consistently with the zero-flux boundary conditions.

Now, proving that the CC-BDF scheme preserves the positivity of the solution is equivalent to showing that $\mathcal{A}^{-1} \geq 0$. To this end, we define D as an $(N+1) \times (N+1)$ diagonal matrix whose diagonal entries are as follows

$$D_{ii} = \frac{1}{\breve{B}_i}, \quad 0 \leq i \leq N,$$

and consider the following lemma where a restriction on the time-step size is required. We remark that with a *convergent matrix* we mean a matrix whose spectral radius is less than one.

Lemma 3. *The matrix $S := I - D\,\mathcal{A}$ is non-negative and convergent, provided that $\delta t < \frac{1}{\gamma}$.*

Proof.

$$S_{ij} = \begin{cases} \frac{\tilde{A}_i}{\breve{B}_i}, & j = i+1, \\ \frac{\tilde{C}_i}{\breve{B}_i}, & j = i-1, \qquad 0 \leq i, j \leq N. \\ 0, & \text{otherwise}, \end{cases}$$

According to (4.16), the matrix S is non-negative since $\tilde{A}_i$, $\breve{B}_i$, $\tilde{C}_i \geq 0$ for $0 \leq i \leq N$.

We show that S is convergent in the sense that $\rho(S)$, the spectral radius of S, is less than one. Based on the theorem of Gerschgorin, $\rho(S) \leq \nu$ where

$$\nu = \max_{0 \leq i \leq N} \sum_{j=0}^{N} |S_{ij}|.$$

It is enough to show that $\nu < 1$. This is proved by contradiction. In fact, the assumption $\nu \geq 1$ means that there exists an i, $0 \leq i \leq N$, such that $\sum_{j=0}^{N} |S_{ij}| \geq 1$. In the following, we prove that in this case $\delta t \geq \frac{1}{\gamma}$, which is contrary to the assumption of the lemma. For $0 < i < N$ we have

$$\begin{aligned}
&\textstyle\sum_{j=0}^{N} |S_{ij}| \geq 1 \Rightarrow \\
&\frac{\tilde{A}_i}{\breve{B}_i} + \frac{\tilde{C}_i}{\breve{B}_i} \geq 1 \Rightarrow \\
&\frac{C^n_{i+\frac{1}{2}} W_i \exp w_i + C^n_{i-\frac{1}{2}} W_{i-1}}{C^n_{i+\frac{1}{2}} W_i + C^n_{i-\frac{1}{2}} W_{i-1} \exp w_{i-1} + \frac{h^2}{\delta t}} \geq 1 \Rightarrow \\
&C^n_{i+\frac{1}{2}} W_i \exp w_i + C^n_{i-\frac{1}{2}} W_{i-1} \geq C^n_{i+\frac{1}{2}} W_i + C^n_{i-\frac{1}{2}} W_{i-1} \exp w_{i-1} + \tfrac{h^2}{\delta t} \Rightarrow \\
&C^n_{i+\frac{1}{2}} W_i (\exp w_i - 1) \geq C^n_{i-\frac{1}{2}} W_{i-1} (\exp w_{i-1} - 1) + \tfrac{h^2}{\delta t} \Rightarrow \\
&C^n_{i+\frac{1}{2}} w_i \geq C^n_{i-\frac{1}{2}} w_{i-1} + \tfrac{h^2}{\delta t} \Rightarrow \\
&B^n_{i+\frac{1}{2}} h \geq B^n_{i-\frac{1}{2}} h + \tfrac{h^2}{\delta t} \Rightarrow \\
&\tfrac{h}{\delta t} \leq B^n_{i+\frac{1}{2}} - B^n_{i-\frac{1}{2}} \leq \gamma h \Rightarrow \\
&\delta t \geq \tfrac{1}{\gamma}.
\end{aligned}$$

We have similar results for $i = 0$ and $i = N$. □

The following lemma is proved in [105].

Lemma 4. *If $\mathcal{A}$ is a real $n \times n$ matrix with non-positive off-diagonal entries, then the following are equivalent:*

1. *$\mathcal{A}$ is nonsingular, and $\mathcal{A}^{-1} \geq 0$.*

2. *The diagonal entries of $\mathcal{A}$ are positive real numbers, and letting D be the diagonal matrix whose diagonal entries are defined as $D_{ii} = \frac{1}{\mathcal{A}_{ii}}$, $1 \leq i \leq n$, then the matrix $S = I - D\mathcal{A}$ is non-negative and convergent.*

Theorem 5. *If $\delta t < \frac{1}{\gamma}$, the discretization scheme (4.14) preserves positivity of the solution of the FP equation.*

Proof. For $\delta t < \frac{1}{\gamma}$, we can employ the Lemma 3 to conclude that $\mathcal{A}^{-1} \geq 0$, based on the Lemma 4. □

4.1.1.2 The Chang-Cooper scheme with second-order time differencing

In the following, we discuss second-order finite difference approximation to the time derivative combined with the CC scheme for approximation in space. Specifically, we consider the second-order backward differentiation formula (BDF2); see, e.g., [39] and references therein.

We have the following discrete FP equation

$$\frac{3f_j^{n+1} - 4f_j^n + f_j^{n-1}}{2\delta t} = D_+ C^n_{j-\frac{1}{2}} D_- f_j^{n+1} + D_+ B^n_{j-\frac{1}{2}} M_\delta f_j^{n+1} + g_j^{n+1}. \qquad (4.19)$$

Summing over j and using the zero-flux boundary conditions, we obtain

$$3\sum_{j=0}^{N} f_j^{n+1} = 4\sum_{j=0}^{N} f_j^{n} - \sum_{j=0}^{N} f_j^{n-1}.$$

By induction and assuming $\sum_{j=0}^{N} f_j^1 = \sum_{j=0}^{N} f_j^0$ and $\sum_{j=0}^{N} f_j^n = \sum_{j=0}^{N} f_j^{n-1}$, we find that

$$\sum_{j=0}^{N} f_j^{n+1} = \sum_{j=0}^{N} f_j^{n}, \quad n \geq 1.$$

Next, we investigate the numerical stability of this CC-BDF2 scheme. We use the following lemma [102].

Lemma 5. *Let (a_k), (b_k), (c_k), and (d_k) be four sequences of non-negative numbers such that the sequence (c_k) is non-decreasing, and*

$$a_k + b_k \leq c_k + \sum_{n=1}^{k-1} d_n a_n, \quad k \geq 2; \quad a_1 + b_1 \leq c_1.$$

Then

$$a_k + b_k \leq c_k \exp\left(\sum_{n=1}^{k-1} d_n\right), \quad k \geq 2.$$

Theorem 6. *If $\delta t \leq \frac{1}{2\gamma}$, then we have the following bound for the solution of the discretization scheme (4.19).*

$$\|f^k\| + \|f^{k-1}\| \leq e^{2\sqrt{2}(k-1)}\left(\|f^0\| + \|f^1\| + 2\delta t \sum_{n=1}^{k-1} \|g^{n+1}\|\right), \qquad k = 2, \cdots, M.$$

Proof. Taking the inner product of equation (4.19) with f^{n+1}, we have

$$\begin{aligned}\left(\frac{3f^{n+1} - 4f^n + f^{n-1}}{2\delta t}, f^{n+1}\right) &= \left(D_+ C^n_{-\frac{1}{2}} D_- f^{n+1}, f^{n+1}\right) + \left(D_+ B^n_{-\frac{1}{2}} M_\delta f^{n+1}, f^{n+1}\right) \\ &+ \left(g^{n+1}, f^{n+1}\right).\end{aligned}$$

First, we note that

$$\begin{aligned}\left(\frac{3f^{n+1} - 4f^n + f^{n-1}}{2\delta t}, f^{n+1}\right) &= \textstyle\sum_{j=0}^{N} \dfrac{h}{2\delta t}(3f_j^{n+1} - 4f_j^n + f_j^{n-1})f_j^{n+1} \\ &= \textstyle\sum_{j=0}^{N} \dfrac{h}{2\delta t}(3|f_j^{n+1}|^2 - 4f_j^n f_j^{n+1} + f_j^{n-1} f_j^{n+1}) \\ &\geq \textstyle\sum_{j=0}^{N} \dfrac{h}{2\delta t}(3|f_j^{n+1}|^2 - 4|f_j^n|^2 - |f_j^{n+1}|^2 \\ &\qquad + \tfrac{1}{2}(f_j^{n-1} + f_j^{n+1})^2 - \tfrac{1}{2}|f_j^{n-1}|^2 - \tfrac{1}{2}|f_j^{n+1}|^2) \\ &\geq \textstyle\sum_{j=0}^{N} \dfrac{h}{2\delta t}(\tfrac{3}{2}|f_j^{n+1}|^2 - 4|f_j^n|^2 - \tfrac{1}{2}|f_j^{n-1}|^2) \\ &= \dfrac{1}{2\delta t}(\tfrac{3}{2}\|f^{n+1}\|^2 - 4\|f^n\|^2 - \tfrac{1}{2}\|f^{n-1}\|^2).\end{aligned}$$

Since

$$\Big(D_+C^n_{-\frac{1}{2}}D_-f^{n+1},f^{n+1}\Big)+\Big(D_+B^n_{-\frac{1}{2}}M_\delta f^{n+1},f^{n+1}\Big)+(g^{n+1},f^{n+1})\leq \tfrac{1}{2}\gamma\|f^{n+1}\|^2+\|g^{n+1}\|\|f^{n+1}\|,$$

we obtain

$$\frac{1}{2\delta t}(\frac{3}{2}\|f^{n+1}\|^2-4\|f^n\|^2-\frac{1}{2}\|f^{n-1}\|^2)\leq\frac{\gamma}{2}\|f^{n+1}\|^2+\|g^{n+1}\|\|f^{n+1}\|,$$

$$\frac{3}{2}\|f^{n+1}\|^2-4\|f^n\|^2-\frac{1}{2}\|f^{n-1}\|^2\leq 2\delta t\|g^{n+1}\|\|f^{n+1}\|+\gamma\delta t\|f^{n+1}\|^2,$$

$$\frac{3}{2}\|f^{n+1}\|^2-4\|f^n\|^2-\frac{1}{2}\|f^{n-1}\|^2\leq 2\delta t^2\|g^{n+1}\|^2+\frac{1}{2}\|f^{n+1}\|^2+\gamma\delta t\|f^{n+1}\|^2,$$

$$\|f^{n+1}\|^2\leq 4\|f^n\|^2+\frac{1}{2}\|f^{n-1}\|^2+2\delta t^2\|g^{n+1}\|^2+\gamma\delta t\|f^{n+1}\|^2,$$

$$(1-\gamma\delta t)\|f^{n+1}\|^2\leq\left(2\|f^n\|+\frac{1}{\sqrt{2}}\|f^{n-1}\|+\sqrt{2}\delta t\|g^{n+1}\|\right)^2.$$

Since $(1-\gamma\delta t)\geq\frac{1}{2}$ we can write

$$\frac{1}{\sqrt{2}}\|f^{n+1}\|-2\|f^n\|-\frac{1}{\sqrt{2}}\|f^{n-1}\|\leq\sqrt{2}\delta t\|g^{n+1}\|.$$

Multiplying with $\sqrt{2}$ and summing over n, we obtain

$$\sum_{n=1}^{k-1}\|f^n\|-\|f^1\|+\|f^k\|-2\sqrt{2}\sum_{n=1}^{k-1}\|f^n\|-\sum_{n=1}^{k-1}\|f^n\|+\|f^{k-1}\|-\|f^0\|\leq 2\delta t\sum_{n=1}^{k-1}\|g^{n+1}\|,$$

$$(-2\sqrt{2})\sum_{n=1}^{k-1}\|f^n\|+\|f^k\|+\|f^{k-1}\|\leq\|f^0\|+\|f^1\|+2\delta t\sum_{n=1}^{k-1}\|g^{n+1}\|,$$

for $2\leq k\leq M$. Hence,

$$\|f^k\|+\|f^{k-1}\|\leq\|f^0\|+\|f^1\|+2\delta t\sum_{n=1}^{k-1}\|g^{n+1}\|+2\sqrt{2}\sum_{n=1}^{k-1}\|f^n\|. \tag{4.20}$$

We apply Lemma 5 to (4.20) with the following setting

$$\begin{aligned}a_k&=\|f^k\|,\quad k\geq 1,\\ b_k&=\|f^{k-1}\|,\quad k\geq 2;\quad b_1=0,\\ c_k=\|f^0\|+\|f^1\|+2\delta t\sum_{n=1}^{k-1}\|g^{n+1}\|,&\quad k\geq 2;\quad c_1=\|f^0\|,\\ d_k&=2\sqrt{2}.\end{aligned}$$

We obtain the following

$$\|f^k\| + \|f^{k-1}\| \leq \mathrm{e}^{2\sqrt{2}(k-1)} \left(\|f^0\| + \|f^1\| + 2\delta t \sum_{n=1}^{k-1} \|g^{n+1}\| \right), \qquad k = 2, \cdots, M.$$

□

Theorem 7. *If $\delta t \leq \frac{1}{2\gamma}$, then the discretization scheme (4.19) converges with error $O(h^2 + \delta t^2)$.*

Proof. By Taylor expansion, we have the following

$$\frac{3f_j^{n+1} - 4f_j^n + f_j^{n-1}}{2\delta t} = \partial_t f_j^n + \delta t \frac{\partial^2}{\partial t^2} f_j^n + \frac{\delta t^2}{6} \frac{\partial^3}{\partial t^3} f_j^n + O(\delta t^3),$$

and

$$\partial_t f_j^n = \partial_t f_j^{n+1} - \delta t \frac{\partial^2}{\partial t^2} f_j^n - \frac{\delta t^2}{2} \frac{\partial^3}{\partial t^3} f_j^n - O(\delta t^3),$$

Therefore, the estimate

$$\partial_t f_j^{n+1} - \frac{3f_j^{n+1} - 4f_j^n + f_j^{n-1}}{2\delta t} = \frac{\delta t^2}{3} \frac{\partial^3}{\partial t^3} f_j^n + O(\delta t^3).$$

Using Equation (4.19), the truncation error can be written as follows

$$\begin{aligned} \phi_j^{n+1} &= [\partial_t f_j^{n+1} - \frac{3f_j^{n+1} - 4f_j^n + f_j^{n-1}}{2\delta t}] + [D_+ C_{j-\frac{1}{2}}^n D_- f_j^{n+1} - \partial_x (C^n \partial_x f)_{x_j}^{t_{n+1}}] \\ &+ [D_+ B_{j-\frac{1}{2}}^n M_\delta f_j^{n+1} - \partial_x (B^n\, f)_{x_j}^{t_{n+1}}]. \end{aligned}$$

Hence the truncation error is $O(h^2 + \delta t^2)$.

Further, it is easily seen that

$$\frac{3e_j^{n+1} - 4e_j^n + e_j^{n-1}}{2\delta t} = D_+ C_{j-\frac{1}{2}}^n D_- e_j^{n+1} + D_+ B_{j-\frac{1}{2}}^n M_\delta e_j^{n+1} + \phi_j^{n+1},$$

hence based on Theorem 6, we have the following

$$\|e^k\| + \|e^{k-1}\| \leq \mathrm{e}^{2\sqrt{2}(k-1)} \left(\|e^0\| + \|e^1\| + 2\delta t \sum_{n=1}^{k-1} \|\phi^{n+1}\| \right).$$

Therefore the difference scheme (4.19) converges with error $O(h^2 + \delta t^2)$. □

The Lax-Richtmyer theorem [66], which states that for consistent numerical approximations stability and convergence are equivalent, shows that the discretization scheme (4.19) is stable.

To investigate the positivity of the CC-BDF2 solution, we consider (4.19) in the semi-discretized form $\partial_t f^n = A_s f^{n+1}$, that is, the FP equation is discretized only in space. The matrix A_s is given by $A_s = \frac{1}{\delta t}(I - \mathcal{A})$, where $\mathcal{A}$ is defined in (4.18). Then the CC-BDF2 scheme is as follows

$$\left(I - \frac{2}{3}\delta t A_s\right) f^{n+1} = \frac{4}{3}f^n - \frac{1}{3}f^{n-1}.$$

Notice that in this formula the presence of a negative factor on the right-hand side prevents us from claiming positivity of the solution for arbitrary $f^0, f^1 \geq 0$. However, as shown in [56], to initialize the BDF2 scheme, we prove that the backward Euler method can be used to compute f^1 such that $2f^1 \geq f^0$. Further, following [19, 73] we prove that under appropriate conditions on the time-step size and on A_s, one has $2f^{n+1} \geq f^n \geq 0$ for $n \geq 1$. As a consequence, we give a sufficient condition that guarantees positivity of the solution obtained with the CC-BDF2 scheme.

We start this discussion with the following lemma.

Lemma 6. *Let A be an $m \times m$ matrix, and $f : \mathbb{R} \to \mathbb{R}^m$. The BDF2 scheme, initialized with the backward Euler scheme, applied to the following equation*

$$\partial_t f = Af, \quad f(0) \geq 0,$$

is positive preserving, provided that $(I - \delta t A)^{-1} \geq 0$, $(I - \frac{2}{3}\delta t A)^{-1} \geq 0$, and $(I + \delta t A) \geq 0$, $(I + 2\delta t A) \geq 0$.

Proof. As discussed in [73], the following backward Euler method

$$\frac{f^1 - f^0}{\delta t} = Af^1$$

can be written as

$$(I - \delta t A)(2f^1 - f^0) = f^0 + \delta t A f^0.$$

Therefore,

$$2f^1 - f^0 = (I - \delta t A)^{-1}((I + \delta t A)f^0),$$

and consequently $2f^1 \geq f^0$, since $(I - \delta t A)^{-1}, (I + \delta t A), f^0 \geq 0$.

The BDF2 method

$$\frac{3f_{\mathrm{j}}^{n+1} - 4f_{\mathrm{j}}^n + f_{\mathrm{j}}^{n-1}}{2\delta t} = Af^{n+1}$$

can be also written in the form

$$(I - \frac{2}{3}\delta t A)(2f^{n+1} - f^n) = \frac{2}{3}(2f^n - f^{n-1}) + \frac{1}{3}(f^n + 2\delta t A f^n).$$

Then it follows that having $2f^n \geq f^{n-1}$, $(I - \dfrac{2}{3}\delta t A) \geq 0$, and $(I + 2\delta t A) \geq 0$, one obtains $2f^{n+1} \geq f^n$. By induction we have $2f^{n+1} \geq f^n$ for $n \geq 1$. □

Theorem 8. *There exists a positive constant q, such that if $\delta t < min\{\frac{1}{\gamma}, \frac{h^2}{2q}\}$, then the discretization scheme (4.19), initialized with the backward Euler scheme, preserves positivity of the solution of the FP equation.*

Proof. The scheme (4.19) is obtained applying the BDF2 scheme to the semi-discretized equation $\partial_t f = A_s f$. From Lemma 4, we know that the condition $\delta t < \frac{1}{\gamma}$ provides $\mathcal{A}^{-1} \geq 0$. Since $\mathcal{A} = I - \delta t A_s$, we can state that if $\delta t < \frac{1}{\gamma}$ then $(I - \delta t A_s)^{-1} \geq 0$. Straightforwardly, we also have $(I - \frac{2}{3}\delta t A_s)^{-1} \geq 0$.

Next, in order to apply Lemma 6, we find a condition such that $(I + \delta t A) \geq 0$ and $(I + 2\delta t A) \geq 0$. For this purpose, notice that

$$(A_s)_{ij} = \begin{cases} \frac{1}{h^2} C^n_{i+\frac{1}{2}} W_j \exp w_i, & j = i+1, \\ -\frac{1}{h^2} Q_i, & j = i, \\ \frac{1}{h^2} C^n_{i-\frac{1}{2}} W_{i-1}, & j = i-1, \\ 0, & otherwise \end{cases} \qquad 0 \leq i, j \leq N,$$

where $Q_i = C^n_{i+\frac{1}{2}} W_i + C^n_{i-\frac{1}{2}} W_{i-1} \exp w_{i-1} \geq 0$. Therefore, the condition $(I + \delta t A) \geq 0$ is equivalent to $(1 - \frac{\delta t}{h^2} Q_i) \geq 0$, and the condition $(I + 2\delta t A) \geq 0$ is equivalent to $(1 - 2\frac{\delta t}{h^2} Q_i) \geq 0$, for all $0 \leq i \leq N$.

Now, we require that $2\frac{\delta t}{h^2} Q_i \leq 1$, that is, $\delta t \leq \frac{h^2}{2Q_i}$. Further, for $0 \leq \delta^n_i \leq 1/2$, we have $Q_i \leq C^n_{i+\frac{1}{2}} + C^n_{i-\frac{1}{2}} \exp w_{i-1}$. Therefore, we choose $q = \max\{C^n_{i+\frac{1}{2}} + C^n_{i-\frac{1}{2}} \exp w_{i-1} : 0 \leq i \leq N\}$. Hence $\delta t \leq \frac{h^2}{2q}$ guarantees $(I + \delta t A) \geq 0$ and $(I + 2\delta t A) \geq 0$.

Taking $\delta t < min\{\frac{1}{\gamma}, \frac{h^2}{2q}\}$, we can apply Lemma 6 to obtain the positivity result. □

4.1.2 Analysis in the multidimensional case

Similar to the one-dimensional case, the CC-BDF difference scheme for the d-dimensional FP equation with diagonal diffusion can be written as follows

$$\frac{f^{n+1}_{\mathrm{j}} - f^n_{\mathrm{j}}}{\delta t} = \sum_{i=1}^{d} \left(D^i_+ C^{i,n}_{\mathrm{j}-1_i/2} D^i_- f^{n+1}_{\mathrm{j}} + D^i_+ B^{i,n}_{\mathrm{j}-1_i/2} M^i_\delta f^{n+1}_{\mathrm{j}} \right) + g^{n+1}_{\mathrm{j}}, \tag{4.21}$$

where

$$D^i_+ C^{i,n}_{\mathrm{j}-1_i/2} D^i_- f^{n+1}_{\mathrm{j}} = \frac{1}{h} \left\{ \frac{1}{h} C^{i,n}_{\mathrm{j}+1_i/2} f^{n+1}_{\mathrm{j}+1_i} - \frac{1}{h} \left(C^{i,n}_{\mathrm{j}+1_i/2} + C^{i,n}_{\mathrm{j}-1_i/2} \right) f^{n+1}_{\mathrm{j}} + \frac{1}{h} C^{i,n}_{\mathrm{j}-1_i/2} f^{n+1}_{\mathrm{j}-1_i} \right\},$$

and

$$\begin{aligned} D^i_+ B^{i,n}_{\mathrm{j}-1_i/2} M^i_\delta f^{n+1}_{\mathrm{j}} &= D^i_+ \left((1 - \delta^{i,n}_{\mathrm{j}-1_i}) B^{i,n}_{\mathrm{j}-1_i/2} f^{n+1}_{\mathrm{j}} + \delta^{i,n}_{\mathrm{j}-1_i} B^{i,n}_{\mathrm{j}-1_i/2} f^{n+1}_{\mathrm{j}-1_i} \right) \\ &= \frac{1}{h} \left\{ (1 - \delta^{i,n}_{\mathrm{j}}) B^{i,n}_{\mathrm{j}+1_i/2} f^{n+1}_{\mathrm{j}+1_i} - (1 - \delta^{i,n}_{\mathrm{j}-1_i}) B^{i,n}_{\mathrm{j}-1_i/2} f^{n+1}_{\mathrm{j}} \right\} \\ &+ \frac{1}{h} \left\{ \delta^{i,n}_{\mathrm{j}} B^{i,n}_{\mathrm{j}+1_i/2} f^{n+1}_{\mathrm{j}} - \delta^{i,n}_{\mathrm{j}-1_i} B^{i,n}_{\mathrm{j}-1_i/2} f^{n+1}_{\mathrm{j}-1_i} \right\}. \end{aligned}$$

Taking the inner product of (4.21) with f^{n+1}, we have

$$\begin{aligned}\left(\frac{f^{n+1}-f^{n}}{\delta t}, f^{n+1}\right) &= \textstyle\sum_{i=1}^{d}\left(D_{+}^{i}C^{i,n}_{-1_i/2}D_{-}^{i}f^{n+1}, f^{n+1}\right)\\ &+ \textstyle\sum_{i=1}^{d}\left(D_{+}^{i}B^{i,n}_{-1_i/2}M_{\delta}^{i}f^{n+1}, f^{n+1}\right) + \left(g^{n+1}, f^{n+1}\right).\end{aligned}$$

With the same argument provided for the one-dimensional case and under similar assumptions on $C^{i}(x,t) > 0$ for $1 \le i \le n$, existence of a constant $\gamma > 0$ such that $\sum_{\mathrm{i}=1}^{n}|B^{i}(x+h,t) - B^{i}(x,t)| \le \gamma h$, and $\delta t \le \frac{1}{2\gamma}$, we may conclude

$$\frac{1}{2\delta t}(\|f^{n+1}\|^{2} - \|f^{n}\|^{2}) \le \frac{1}{2}\gamma\|f^{n+1}\|^{2} + \|g^{n+1}\|\|f^{n+1}\|,$$

which leads to the following bound for the solution

$$\|f^{k}\| \le 2^{k/2}\|f^{0}\| + \delta t\sum_{n=0}^{k-1}2^{\frac{k-n+1}{2}}\|g^{n+1}\|,$$

and consequently the convergence order, $O(h^{2}+\delta t)$.

In a similar way, it is proven that the CC-BDF2 scheme has order of convergence $O(h^{2}+\delta t^{2})$.

To investigate positivity of the solution, we consider the CC-BDF scheme in the following form

$$\begin{aligned}\frac{f_{\mathrm{j}}^{n+1}-f_{\mathrm{j}}^{n}}{\delta t} = \textstyle\sum_{i=1}^{d}\frac{1}{h^{2}} \quad \{ \quad & C^{i,n}_{\mathrm{j}+1_i/2}W^{i,n}_{\mathrm{j}}\exp w^{i,n}_{\mathrm{j}}f^{n+1}_{\mathrm{j}+1_i}\\ - \quad & (C^{i,n}_{\mathrm{j}+1_i/2}W^{i,n}_{\mathrm{j}} + C^{i,n}_{\mathrm{j}-1_i/2}W^{i,n}_{\mathrm{j}-1_i}\exp w^{i,n}_{\mathrm{j}-1_i})f^{n+1}_{\mathrm{j}}\\ + \quad & C^{i,n}_{\mathrm{j}-1_i/2}W^{i,n}_{\mathrm{j}-1_i}f^{n+1}_{\mathrm{j}-1_i}\}\end{aligned}$$

where $W^{i,n}_{\mathrm{j}} = w^{i,n}_{\mathrm{j}}/(\exp w^{i,n}_{\mathrm{j}} - 1)$. This can be written in the form of a system of linear equations composed of $(N+1)^{d}$ unknowns f^{n+1}_{j}, $|\mathrm{j}| \le dN$. To this end, we give an order to the mesh points and associate integer j to the mesh point x_{j} with the rule $j = \sum_{i=1}^{d}(N+1)^{i-1}j_{i}$. The $(N+1)^{d}$-dimensional vector f^{n} having the given value f^{n}_{j} as the j-th element, $0 \le j \le (N+1)^{d}-1$, constitutes the right-hand side of the system. The matrix of the coefficients $\mathcal{A}$ is an $(N+1)^{d} \times (N+1)^{d}$ matrix with the elements

$$\mathcal{A}_{pq} = \begin{cases} -\alpha, & q = p + (N+1)^{i-1},\ i = 1,\cdots,d,\\ \beta, & q = p,\\ -\gamma, & q = p - (N+1)^{i-1},\ i = 1,\cdots,d,\\ 0, & \text{otherwise,}\end{cases} \qquad 0 \le p,q \le (N+1)^{d}-1,$$

where

$$\alpha = \frac{\delta t}{h^{2}}C^{i,n}_{\mathrm{j}+1_i/2}W^{i,n}_{\mathrm{j}}\exp w^{i,n}_{\mathrm{j}},$$

$$\beta = 1 + \sum_{i=1}^{d} \frac{\delta t}{h^2}(C^{i,n}_{\mathrm{j}+1_i/2} W^{i,n}_{\mathrm{j}} + C^{i,n}_{\mathrm{j}-1_i/2} W^{i,n}_{\mathrm{j}-1_i} \exp w^{i,n}_{\mathrm{j}-1_i}),$$

$$\gamma = \frac{\delta t}{h^2} C^{i,n}_{\mathrm{j}-1_i/2} W^{i,n}_{\mathrm{j}-1_i},$$

and j is the corresponding multi-index to the integer p.

The same argument presented for the one-dimensional case shows that $\mathcal{A}^{-1} \geq 0$ provided $\delta t < \frac{1}{\gamma}$. Therefore, under this condition, the CC-BDF scheme produces positive solutions starting from positive initial conditions.

As in the one-dimensional case, following the reasoning of Theorem 8, we can prove positivity of the CC-BDF2 scheme.

4.1.3 Numerical experiments

In this section, we present results of numerical experiments to validate our theoretical findings. We consider the FP problem corresponding to the Ornstein-Uhlenbeck process, where $B(x,t) = \gamma_0 x$, and $C(x,t) = \sigma^2$, and γ_0 and σ are two constants of the stochastic process. Specifically, we choose $\gamma_0 = 1$ and $\sigma = 1$. Further, we take $T = 1$ and $\Omega = (0, L)$.

We assume zero-flux boundary conditions at the boundary of $\Omega = (0, L)$. Choosing $g(x,t) = (L-x)(2x-L)/\exp((x-L/2)^2+t)$ and $f_0(x) = 1/\exp((x-L/2)^2)$, we have the exact solution $f_{exact}(x,t) = 1/\exp((x-L/2)^2+t)$ whose corresponding flux tends to zero at the boundary of the domain $\Omega = (0, L)$. For $L = 10$ the flux is of order 10^{-10}.

The size of the solution error is evaluated based on the following L^2-norm. We have

$$\|v\|^2_{L^2_{h,\delta t}(Q)} = h\delta t \sum_{n=1}^{M} \sum_{j=0}^{N} |v^n_j|^2,$$

where v is a space-time grid function.

In Table 4.1, we report results of experiments that evaluate the accuracy of the CC-BDF numerical solution. We see that the resulting order of convergence is $O(h^2 + \delta t)$.

N	M	$\|f - f_{exact}\|_{L^2_{h,\delta t}(Q)}$
50	50	1.34e-2
100	200	3.5e-3
200	800	8.8097e-4

Table 4.1: Convergence of the CC-BDF scheme.

We use the same setting to validate the CC-BDF2 scheme. The corresponding results are given in Table 4.2, that confirm second-order convergence in space and time. With the CC-BDF2 scheme, two cases are considered concerning initial conditions. We denote with f_α the numerical solution obtained in the case when both f^0 and

f^1 are exactly known. Further, we denote with f_β the numerical solution obtained in the case where only f^0 is exactly given, and f^1 is numerically approximated by a first-order backward Euler method with time-step size δt^2. The norms of the solution errors resulting from these two approaches are give in Table 4.2. We see that second-order convergence in space and time is obtained, that is the convergence is of order $O(h^2 + \delta t^2)$.

N	M	$\|f_\alpha - f_{exact}\|_{L^2_{h,\delta t}(Q)}$	$\|f_\beta - f_{exact}\|_{L^2_{h,\delta t}(Q)}$
50	50	1.46e-2	1.57e-2
100	100	3.8e-3	4.0e-3
200	200	9.5491e-4	1.0e-3

Table 4.2: Convergence of the CC-BDF2 scheme with two different initializations.

By numerical inspection, we find that the proposed schemes preserve positiveness and are conservative.

4.2 The Fokker-Planck equation for PDP processes

In this section, we consider another class of FP equations, where the PDF corresponds to a piecewise deterministic process. We recall the PDP model introduced in Chapter 3 which is a first-order system of ordinary differential equations, where the driving dynamics-function is chosen by a renewal process. The d-components state function $X(t)$, $X : [t_0, \infty) \to \Omega$, $\Omega \subseteq \mathbb{R}^d$, satisfies the differential equation

$$\frac{d}{dt}X(t) = A_{\mathscr{S}(t)}\left(X(t)\right), \quad t \in [t_0, \infty), \tag{4.22}$$

where $\mathscr{S}(t) : [t_0, \infty[\to \mathbb{S}$ is a Markov process with discrete states $\mathbb{S} = \{1, \ldots, S\}$. Let us denote with f_s the PDF corresponding to the state s. The time evolution of these PDFs is governed by the following FP hyperbolic system,

$$\partial_t f_s(x,t) + \partial_x(A_s(x) f_s(x,t)) = \sum_{j=1}^{S} \mathcal{Q}_{sj} f_j(x,t), \qquad s \in \mathbb{S}, \tag{4.23}$$

where $\mathcal{Q}_{sj}$, $s, j \in \mathbb{S}$, are the components of the transition matrix $\mathcal{Q}$.

We present finite difference discretization for the FP model (4.23). The discretization scheme is first-order that guarantees positivity and conservativeness of the numerical PDF solution. We focus on a particular PDP process. This choice is motivated by the wish to provide a detailed discussion and implementation for a specific problem. We consider the case of a dissipative process subject to dichotomic noise; see, e.g., [4, 5, 89]. The finite difference discretization for the corresponding FP equation is presented from [8].

4.2.1 A PDP process with dichotomic noise

Let $X(t)$ be a process whose evolution is described by the following equation

$$\frac{dX(t)}{dt} = -\gamma\, X + W\, \xi(t), \tag{4.24}$$

where the noised input $\xi(t)$ represents a dichotomic noise (or random telegraph signal), that takes values ± 1, with exponential statistics of the switching time given by the PDF $\mu\, \exp(-\mu t)$. The solution to (4.24) is composed of pieces of increasing and decreasing exponentials, but the whole process $X(t)$ is not deterministic. In fact, it represents a random sample path in a probability space. For simplicity, we take $\gamma = 1$ and $W = 1$. Therefore, we have the following dynamics

$$A_1(x) = 1 - x, \qquad A_2(x) = -(1 + x). \tag{4.25}$$

For the purpose of illustration, we depict in Figure 4.1 the characteristics of the hyperbolic FP system (4.23) corresponding to the setting (4.25). There are two families of curves, one for each state of the system. It is clear that if the initial non-zero PDF data is contained in the interval $(-1, 1)$, then the PDF will never escape from this interval during the time evolution. On the other hand, the PDF outside $(-1, 1)$ remains equal to zero. Within $(-1, 1)$, the PDF can switch between characteristics belonging to the two different families, but it will never cross the points $x = \pm 1$. We name the interval $(-1, 1)$ the invariant set, while the set $\mathbb{R}\backslash[-1, 1]$ is named transient, since an initial PDF located in this region will soon be transported and trapped in the invariant set.

For our special setting, the system of equations (4.23) become

$$\begin{aligned} \partial_t f_1(x,t) + \partial_x\left((1-x)\, f_1(x,t)\right) &= -\mu\, f_1(x,t) + \mu\, f_2(x,t) & (4.26)\\ \partial_t f_2(x,t) - \partial_x\left((1+x)\, f_2(x,t)\right) &= +\mu\, f_1(x,t) - \mu\, f_2(x,t), & (4.27) \end{aligned}$$

Further, we assume appropriate initial conditions

$$f_s(x, 0) = f_s^0(x), \qquad s = 1, 2, \tag{4.28}$$

where $f_s^0(x) \geq 0$, $\sum_{s=1}^{2} \int_\Omega f_s^0(x) = 1$.

Combining (4.26) and (4.27), and using integration by parts, we obtain

$$\partial_t \left(\int_\Omega f_1(x,t)\, dx + \int_\Omega f_2(x,t)\, dx \right) = 0.$$

This proves conservativeness of the FP system. In particular, we have

$$\int_\Omega f_1(x,t)\, dx + \int_\Omega f_2(x,t)\, dx = \int_\Omega f_1^0(x) + \int_\Omega f_2^0(x) = 1. \tag{4.29}$$

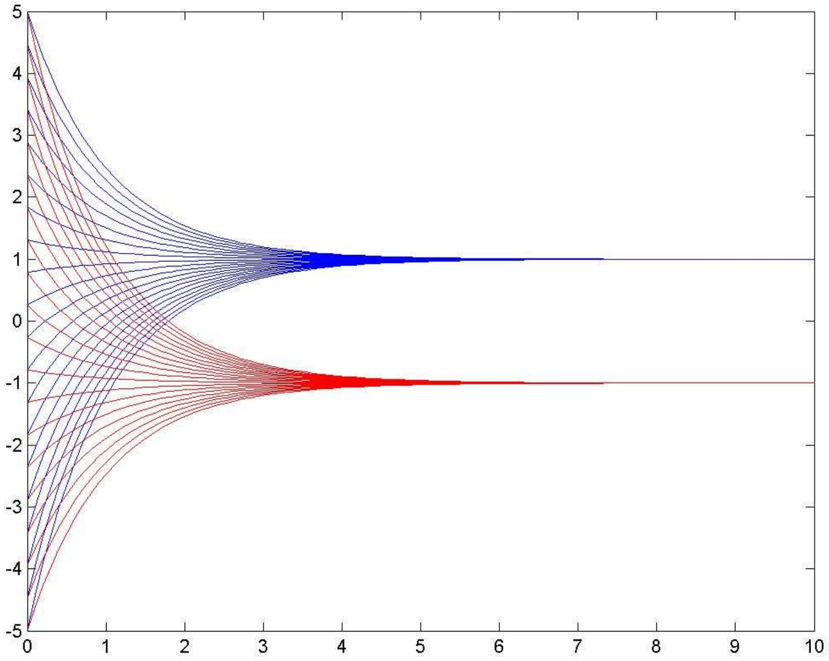


Figure 4.1: Characteristics of the FP equation for a PDP process with dichotomic noise.

4.2.2 Discretization of the FP system

Consider the dichotomic FP problem in the time interval $(0, T)$ and a bounded spatial domain $\Omega = (-L, L)$. We choose L sufficiently large to include the invariant set and the space-time support of the FP solution.

For the time discretization, we define the time-step size $\delta t = T/M$, in which M is a positive integer, and $I_{\delta t} = \{t_n = n\,\delta t, n = 0, 1, \ldots, M\}$ is the time mesh. Moreover, we consider a uniform mesh on the state space with mesh size $h = 2L/N$, and the mesh-point coordinates are denoted with $x_j = jh - L, j = 0, \ldots, N$. We have the following mesh

$$\Omega_h = \{x \in \mathbb{R} : x_j = j\,h,\; j \in \mathbb{Z}\} \cap \Omega.$$

For grid functions w and v defined on $\Omega_h \times I_{\delta t}$, with values $w_j^n = w(x_j, t_n)$, we introduce the discrete L^2-scalar product

$$(w^n, v^n)_h = h \sum_{j=1}^{N-1} w_j^n v_j^n,$$

with associated L^2-norm $\|v^n\| = (v^n, v^n)^{1/2}$. Further, we introduce the following inner product

$$\langle w, v \rangle_{h,\delta t} = \delta t \sum_{n=0}^{M-1} (w^n, v^n)_h.$$

Here and below, we denote the approximation to $f(x_j, t_n)$ with f_j^n.

Next, we consider the following explicit discretization of the time derivative

$$D_t f_j^n = \frac{1}{\delta t} \left(f_j^{n+1} - f_j^n \right), \tag{4.30}$$

and the first-order forward and backward space derivatives are given by

$$D_x^+ f_j^n = \frac{1}{h} \left(f_{j+1}^n - f_j^n \right), \qquad D_x^- f_j^n = \frac{1}{h} \left(f_j^n - f_{j-1}^n \right).$$

For ease of notation, let us denote with $f = f_1$ and $g = f_2$. According to upwind discretization, the discrete FP equations are as follows

$$D_t f_j^n + D_x^- \left((1 - x_j)\, f_j^n \right)^+ + D_x^+ \left((1 - x_j)\, f_j^n \right)^- \;=\; -\mu\, f_j^n + \mu\, g_j^n \tag{4.31}$$

$$D_t g_j^n - D_x^+ \left((1 + x_j)\, g_j^n \right)^+ - D_x^- \left((1 + x_j)\, g_j^n \right)^- \;=\; +\mu\, f_j^n - \mu\, g_j^n, \tag{4.32}$$

where the symbols $(f)^+$ and $(f)^-$, denote $\max(f, 0)$ and $\min(f, 0)$, respectively, and we assume the boundary conditions $f_0^n = 0$ and $g_N^n = 0$ for $n = 0, \ldots, M$. Further, we have the following initial conditions

$$f_j^0 = f_{0,j}, \qquad g_j^0 = g_{0,j}. \tag{4.33}$$

We have the following theorem.

Theorem 9. *The discretization scheme (4.31)-(4.32) is stable, positivity preserving, conservative, and first-order accurate with respect to the solution of Eqs. (4.26)-(4.27), provided that the following Courant-Friedrichs-Lèwy (CFL) condition on the time step size is satisfied*

$$\delta t \leq \min \left\{ \frac{h}{2 + u_1 + \mu h}, \frac{h}{2 + u_2 + \mu h} \right\}. \tag{4.34}$$

Proof. Let us define the grid functions

$$b_1(x_j) = (1 - x_j) \frac{\delta t}{h}, \qquad b_2(x_j) = -(1 + x_j) \frac{\delta t}{h}.$$

The discrete FP system (4.31)-(4.32) becomes

$$\begin{aligned} f_j^{n+1} &= f_j^n (1 - |b_1(x_j)|) + (b_1(x_{j-1}) f_{j-1}^n)^+ - (b_1(x_{j+1}) f_{j+1}^n)^- \\ &\quad - \delta t \mu\, f_j^n + \delta t \mu\, g_j^n \end{aligned} \tag{4.35}$$

$$\begin{aligned} g_j^{n+1} &= g_j^n (1 - |b_2(x_j)|) + (b_2(x_{j+1}) g_{j+1}^n)^+ - (b_2(x_{j-1}) g_{j-1}^n)^- \\ &\quad + \delta t \mu\, f_j^n - \delta t \mu\, g_j^n. \end{aligned} \tag{4.36}$$

Further, we have

$$
\begin{aligned}
f_j^{n+1} + g_j^{n+1} \;=\;& f_j^n(1 - |b_1(x_j)|) + (b_1(x_{j-1}))^+ f_{j-1}^n - (b_1(x_{j+1}))^- f_{j+1}^n \\
& + g_j^n(1 - |b_2(x_j)|) + (b_2(x_{j+1}))^+ g_{j+1}^n - (b_2(x_{j-1}))^- g_{j-1}^n.
\end{aligned} \tag{4.37}
$$

By inspection of (4.35)-(4.36) and (4.37), we conclude that positivity of f_j^n, g_j^n and of $f_j^n + g_j^n$, $n = 1, \dots, M$ and $j = 0, \dots, N$, is guaranteed if the following requirements are satisfied:

$$\max_{x \in \Omega} (|b_1(x)|, |b_2(x)|) \leq 1$$

and

$$\min_j (1 - b_1(x_j)) - \delta t \mu \geq 0, \qquad \min_j (1 - b_2(x_j)) - \delta t \mu \geq 0.$$

Notice that we assume that $f_j^0 + g_j^0 \geq 0$ for $j = 0, \dots, N$. Now, it is immediate to see that these two conditions hold if (4.34) is true; see also [4].

Next, to prove conservativeness, we take the sum of (4.37) on all interior grid points. We obtain

$$\sum_{j=1}^{N-1} \left(f_j^{n+1} + g_j^{n+1}\right) = \sum_{j=1}^{N-1} \left(f_j^n + g_j^n\right) - \delta t\, f_{N-1}^n - \delta t\, g_1^n, \tag{4.38}$$

where we use the fact $b_1(x_{M-1}) = \delta t$ and $b_2(x_1) = \delta t$. Notice that in the case of a compactly supported solution such that $f_{N-1}^n = 0$ and $g_1^n = 0$, our first order scheme is conservative and positive preserving. In addition, the same result proves that the scheme is stable.

Next, we discuss the accuracy of the scheme (4.31)-(4.32). We assume sufficient regularity of the data such that the solution of the continuous problem is twice continuously differentiable in space and time. Denote with ϕ and ψ the local truncation error for the discretization of (4.31) and (4.32), respectively. It is a standard calculation [4] to show that the following holds

$$|\phi_j^n| \leq C\,(\delta t + h), \qquad |\psi_j^n| \leq C\,(\delta t + h), \tag{4.39}$$

where C depends on the maximum of the second derivatives of f and g in the space time domain. Therefore, it is a consequence of the Lax-Richtmyer equivalence theorem [83] that the estimate (4.39) and the stability of the scheme prove that the scheme is first-order accurate. □

As a numerical experiment, we assume that initial distribution of the PDP system is given by two Gaussian distributions, defined in $\Omega = (-2, 2)$, centered in $x = 0$ and variance $\sigma^2 = 10^{-2}$. The transition rate of the underlying Markov process is $\mu = 2$. See Figure 4.2 for a snapshot of the FP solution at $t = 10$.

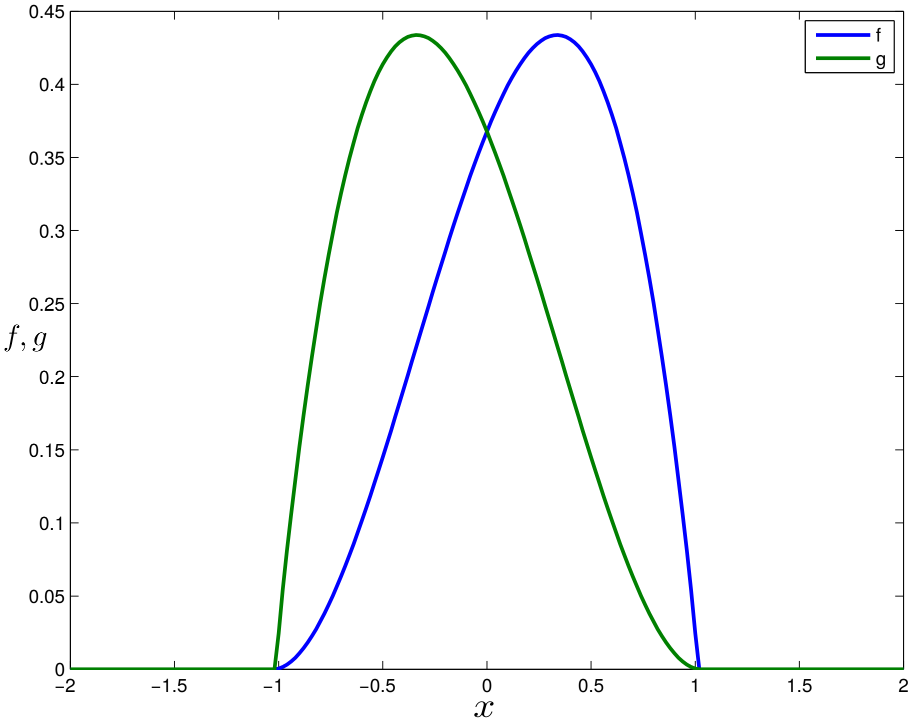


Figure 4.2: The PDF functions at time $t = 10$ resulting from the FP evolution.

Chapter 5

Hermite spectral discretization of FP equations

5.1 Spectral methods

Spectral methods involve seeking the solution to a differential equation in terms of a series of known, smooth functions which are called basis functions. They involve representing the solution to a problem as truncated series of known functions of the independent variables. Along with extensive applications of Legendre and Chebyshev spectral methods for bounded domains, considerable progress has been made recently in spectral methods for unbounded domains. Among these methods, a direct and commonly used approach is based on certain orthogonal approximations on infinite intervals, in particular the Hermite and Laguerre spectral methods [52, 97].

Spectral methods for solving PDEs on unbounded domains can be essentially classified into four approaches [97]:
(i) Domain truncation: truncate unbounded domains to bounded domains and solve the PDEs on bounded domains supplemented with artificial or transparent boundary conditions;
(ii) Approximation by classical orthogonal systems on unbounded domains, e.g., Laguerre and Hermite polynomials/functions;
(iii) Approximation by other, non-classical orthogonal systems, or by mapped orthogonal systems, e.g., image of classical Jacobi polynomials through a suitable mapping;
(iv) Mapping: map unbounded domains to bounded domains and use standard spectral methods to solve the mapped PDEs in the bounded domains.

In general, the domain truncation approach is only a viable option for problems with rapidly (exponentially) decaying solutions or when accurate non-reflecting or exact boundary conditions are available at the truncated boundary. On the other hand, with proper choices of mappings and/or scaling parameters, the other three approaches can all be effectively applied to a variety of problems with rapid or slow decaying (or even growing) solutions. We note that the last two approaches are mathematically equivalent, but their computational implementations are different. More precisely, the last approach involves solving the mapped PDEs (which are often cumbersome to deal with) using classical Jacobi polynomials while the approach (iii) solves the original PDE using the mapped Jacobi polynomials. The main advantage of the approach (iv) is that it can be implemented and analyzed using standard procedures and approximation results, but its main disadvantage is that the transformed equation is usually very complicated which, in many cases, makes its implementation and analysis unusually cumbersome [95].

While spectral methods have been used for solving PDEs on unbounded domains for over thirty years, and there have been several isolated efforts in the early years on the error analysis of these methods, it is only in the last ten years or so that the basic approximation properties of these orthogonal systems, and their applications to PDEs, were systematically studied [95].

5.2 Hermite approximation space

5.2.1 Hermite polynomials

The Hermite polynomials, denoted by $H_n(x)$, are the eigenfunctions of the Sturm-Liouville problem:

$$e^{x^2}\frac{d}{dx}\left(e^{-x^2}\frac{d}{dx}u(x)\right)+\lambda u(x)=0,\quad x\in\mathbb{R},$$

with the eigenvalue $\lambda_n = 2n$ grows linearly with respect to n.

The Hermite polynomials are orthogonal with respect to the weight $w(x) = e^{-x^2}$, i.e.,

$$\int_{-\infty}^{\infty} H_n(x)H_m(x)e^{-x^2}dx=\gamma_n\delta_{n,m},$$

where $\gamma_n = \sqrt{\pi}2^n n!$ and $\delta_{n,m}$ is the Kronecker delta. Note that the constant γ_n grows exponentially as n increases, so it is necessary to normalize this factor in actual computations. The three-term recurrence formula reads

$$H_{n+1}(x)=2xH_n(x)-2nH_{n-1}(x),\quad n\geq 1,$$

and the first few members are

$$\begin{aligned}
H_0(x) &= 1,\\
H_1(x) &= 2x,\\
H_2(x) &= 4x^2-2,\\
H_3(x) &= 8x^3-12x,\\
H_4(x) &= 16x^4-48x^2+12.
\end{aligned}$$

One verifies by induction that the leading coefficient of $H_n(x)$ is 2^n. The Hermite polynomials have a close connection with the generalized Laguerre polynomials [97]:

$$\begin{aligned} H_{2n}(x) &= (-1)^n 2^{2n} n! \mathcal{L}_n^{-1/2}(x^2), \\ H_{2n+1}(x) &= (-1)^n 2^{2n+1} n! x \mathcal{L}_n^{1/2}(x^2). \end{aligned}$$

Hence, $H_n(x)$ is odd (resp. even) for n odd (resp. even), that is,

$$H_n(-x) = (-1)^n H_n(x).$$

Moreover,

$$H_{2n}(0) = (-1)^n \frac{(2n)!}{n!}, \quad H_{2n+1}(0) = 0.$$

We also have the orthogonality

$$\int_{-\infty}^{\infty} \frac{d}{dx} H_n(x) \frac{d}{dx} H_m(x) e^{-x^2} dx = \lambda_n \gamma_n \delta_{n,m},$$

The Hermite polynomials are generally not suitable in practice due to their asymptotic behavior at infinities [95]:

$$\begin{aligned} H_n(x) &\sim \frac{\Gamma(n+1)}{\Gamma(n/2+1)} e^{x^2/2} \cos\left(\sqrt{2n+1}x - \frac{n\pi}{2}\right) \\ &\sim n^{n/2} e^{x^2/2} \cos\left(\sqrt{2n+1}x - \frac{n\pi}{2}\right). \end{aligned}$$

5.2.2 Convergence of Hermite expansions and the Gibbs phenomenon

Since in some of our FP problems we encounter singularities in the equilibrium solutions, we present here the related discussion from [49], which explains the Gibbs phenomenon occurring at the singular points.

We first consider the expansion of a function $f(x)$ in terms of the eigenfunctions ϕ_n of a Sturm-Liouville problem. The eigenfunction $\phi_n(x)$ is a nonzero solution to

$$\frac{d}{dx} p(x) \frac{d\phi_n}{dx} + (\lambda_n w(x) - q(x))\phi_n(x) = 0 \tag{5.1}$$

satisfying homogeneous boundary conditions in the interval $[a, b]$. To be specific, we assume the boundary conditions $\phi_n(a) = \phi_n(b) = 0$, although the analysis applies more generally. We assume that $p(x) \geq 0$, $w(x) > 0$, $q(x) \geq 0$ for $a \leq x \leq b$. We will also assume that the eigenfunctions are normalized so that they satisfy

$$\int_a^b w(x)\phi_n(x)\phi_m(x) dx = \delta_{nm}, \tag{5.2}$$

and that they form a complete set; the latter property follows if $\lambda_n \to \infty$ as $n \to \infty$. The requirement that $\lambda_n \to \infty$ follows heuristically as follows: (5.1) suggests that

$\phi_n(x)$ has a typical spatial scale of $1/\sqrt{\lambda_n}$, so the requirement that arbitrary $f(x)$ be expansible in terms of ϕ_n implies that λ_n must grow unboundedly with n.

We wish to estimate the rate of convergence of the eigenfunction expansion

$$f(x) = \sum_{n=1}^{\infty} a_n \phi_n(x). \tag{5.3}$$

Using the orthonormality relation (5.2), the L_2-error after N terms is

$$\left[\int_a^b |f(x) - \sum_{n=1}^{N} a_n \phi_n(x)|^2 w(x) dx\right]^{1/2} = \left[\sum_{n=N+1}^{\infty} a_n^2\right]^{1/2}.$$

Thus, the L_2-error may be estimated by calculating the rate of decrease of a_n as $n \to \infty$. Orthonormality of ϕ_n implies that

$$a_n = \int_a^b f(x)\phi_n(x)w(x)dx. \tag{5.4}$$

Substituting $w(x)\phi_n(x)$ from the Sturm-Liouville equation (5.1) gives

$$a_n = \frac{1}{\lambda_n}\int_a^b \left(-\frac{d}{dx}p(x)\frac{d\phi_n}{dx} + q(x)\phi_n\right) f(x)dx.$$

Integrating twice by parts, we obtain

$$a_n = \frac{1}{\lambda_n}p(x)[\phi_n(x)f'(x) - \phi_n'(x)f(x)]|_{x=a}^{b} + \frac{1}{\lambda_n}\int_a^b h(x)\phi_n(x)w(x)dx, \tag{5.5}$$

where

$$h(x) = [-\frac{d}{dx}p(x)\frac{df}{dx} + q(x)f(x)]/w(x). \tag{5.6}$$

This integration by parts is justified if f is twice differentiate and h is square integrable with respect to w. Under these conditions and recalling that $\phi_n(a) = \phi_n(b) = 0$, we obtain

$$a_n = \frac{1}{\lambda_n}[p(a)\phi_n'(x)f(a) - p(b)\phi_n'(b)f(b)] + O(\frac{1}{\lambda_n})$$

as $n \to \infty$, since $|\int_a^b h\phi_n w dx|^2 \leqq \int_a^b h^2 w dx \int_a^b \phi_n^2 w dx = O(1)$ as $n \to \infty$.

To proceed further we must distinguish between nonsingular and singular Sturm-Liouville problems. A problem is nonsingular if $p(x) > 0$ and $w(x) > 0$ throughout $a \leq x \leq b$. The important conclusion from (5.5)-(5.6) is that if the Sturm-Liouville problem is nonsingular and if $f(a)$ or $f(b)$ is nonzero then

$$a_n \sim \frac{1}{\lambda_n}[p(a)\phi_n'(x)f(a) - p(b)\phi_n'(b)f(b)], \quad n \to \infty. \tag{5.7}$$

Notice that if $\phi_n'(a) = 0$, then $\phi_n(x) \equiv 0$ since (5.1) is a second-order differential equation and $p(x) \neq 0$. It is well known that the asymptotic behavior of the eigenvalues and eigenfunctions of a nonsingular Sturm-Liouville problem are given by

$$\lambda_n \sim \left[n\pi \Big/ \int_a^b \sqrt{\frac{w}{p}} dx \right], \quad n \to \infty, \tag{5.8}$$

$$\phi_n(x) \sim A_n \sin\left(\sqrt{\lambda_n} \int_a^x \sqrt{\frac{w}{p}} dx \right), \quad n \to \infty. \tag{5.9}$$

Using (5.8)-(5.9) in (5.7), we find that a_n behaves like $1/n$ as $n \to \infty$ if either $f(a) \neq 0$ or $f(b) \neq 0$. This behavior of a_n leads to the Gibbs phenomenon in the expansion (5.3) near those boundary points at which $f(a)$ or $f(b) \neq 0$. If $f(a) = f(b) = 0$, then $a_n \ll 1/n$ as $n \to \infty$. However, a further integration by parts in (5.5) shows that if the Sturm-Liouville problem is nonsingular and if $h(a)$ or $h(b) \neq 0$, then a_n behaves like $1/n^3$ as $n \to \infty$. In general, unless $f(x)$ satisfies an infinite number of very special conditions at $x = a$ and $x = b$, then a_n decays algebraically as $n \to \infty$. These results on algebraic decay of errors in expansions based on nonsingular second-order eigenvalue problems generalize to higher-order eigenvalue problems.

If $p(a) = 0$ in (5.7) then it is not necessary to require that $f(a) = 0$ to achieve $a_n \ll \phi_n'/\lambda_n$ as $n \to \infty$. For this reason, expansions based on eigenfunctions of a Sturm-Liouville problem that is singular at $x = a$ do not normally exhibit the Gibbs phenomenon at $x = a$. Furthermore, if the argument that led to (5.7) can be repeated on $h(x)$ given by (5.6) (this is possible if p/w, p'/w, and q/w are bounded and all derivatives of f are square integrable with respect to w) then the boundary contribution to a_n from $x = a$ is smaller than ϕ_n'/λ_n^2 as $n \to \infty$. If there are also no boundary contributions from $x = b$ when the operations leading to (5.7) are repeated indefinitely (which is true if $p(b) = 0$), then a_n decreases more rapidly than any power of $1/\lambda_n$ as $n \to \infty$.

The important conclusion is that eigenfunction expansions based on Sturm-Liouville problems that are singular at $x = a$ and at $x = b$ converge at a rate governed by the smoothness of the function being expanded not by any special boundary conditions satisfied by the function.

Hermite polynomials satisfy (5.1) with $p = e^{-x^2}$, $q(x) = 0$, $w(x) = e^{-x^2}$ for $-\infty < x < \infty$, $\phi_n(x)e^{-x^2/2}$ bounded as $|x| \to \infty$. The Hermite polynomial $H_n(x)$ of degree n is associated with the eigenvalue $\lambda_n = 2n$. If $f(x)$ and all its derivatives satisfy

$$f(x) = O(e^{\alpha x^2}), \qquad |x| \to \infty,$$

for some $\alpha < \frac{1}{2}$, then the Hermite expansion

$$f(x) = \sum_{n=0}^{\infty} a_n H_n(x)$$

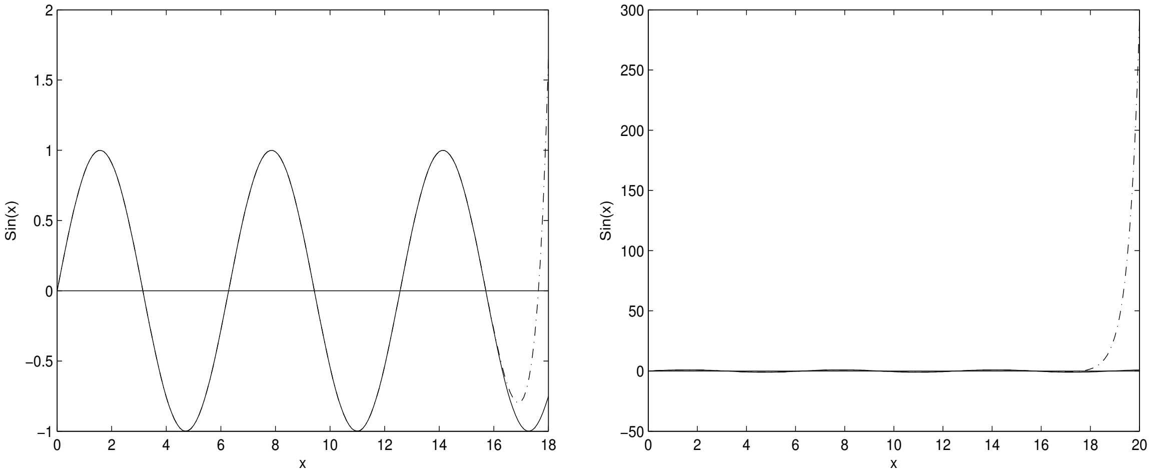


Figure 5.1: Exact representation (solid line) and Hermite polynomial approximation (dash-dot line) of $\sin(x)$. The Hermite polynomial expansion is truncated after $N = 20$ terms. The spatial domain is $[0, 18]$ in the left figure, and $[0, 20]$ in the right figure.

converges faster than algebraically as the number of terms $N \to \infty$. This is proved by retracting the steps leading from (5.4) to (5.7).

To study the rate of convergence of Hermite series, we consider the expansion of $\sin x$ as follows.

$$\sin x = \sum_{n=0}^{\infty} \frac{1}{2^{2n+1}(2n+1)!} H_{2n+1}(x). \tag{5.10}$$

Since the asymptotic behavior of $H_n(x)$ is given by

$$H_n(x) \sim e^{x^2/2} \frac{n!}{\frac{1}{2}n!} \cos(\sqrt{2n+1}x - \frac{1}{2}n\pi)$$

as $n \to \infty$ for fixed x, it follows that the error after N terms of (5.10) goes to zero rapidly at x only if $N \gtrsim x^2/\log x$. This result is not good; to resolve m wavelengths of $\sin x$ requires nearly m^2 Hermite polynomials; see Figure 5.1 and Figure 5.2. By expanding in the series $\sum a_n H_n(x) e^{-\alpha x^2}$ and optimizing the choice of α, it is possible to reduce the number of required Hermite polynomials to about $\frac{5}{2}\pi \approxeq 7.85$ per wavelength.

5.2.3 Hermite functions

Since the Hermite polynomials are not suitable for our FP problems due to their asymptotic behavior at infinities, we shall consider the so called Hermite functions. Hermite functions are defined as follows

$$\tilde{\mathrm{H}}_n(x) = \frac{1}{\sqrt{2^n n!}} H_n(\alpha x) w_\alpha^{-1}(x), \quad \alpha > 0, \;\; n \geq 0,$$

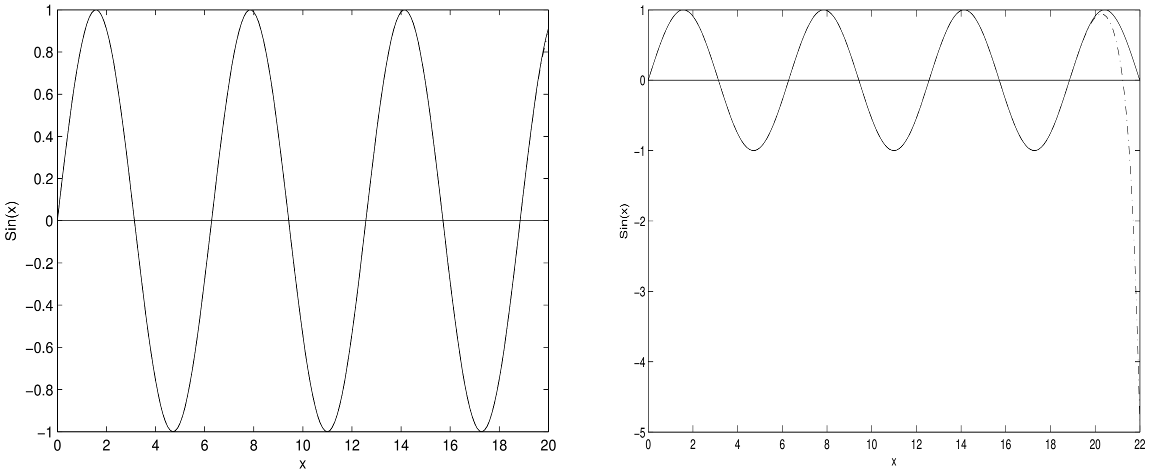


Figure 5.2: Exact representation (solid line) and Hermite polynomial approximation (dash-dot line) of $\sin(x)$. The Hermite polynomial expansion is truncated after $N = 25$ terms. The spatial domain is $[0, 20]$ in the left figure, and $[0, 22]$ in the right figure.

where $w_\alpha(x) = \exp(\alpha^2 x^2)$ is a weight function and H_n is the Hermite polynomial of degree n given by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2}).$$

The function $\tilde{\mathrm{H}}_n(x)$ is the n-th eigenfunction of the following singular Liouville problem:

$$\frac{d}{dx}\left(e^{-\alpha^2 x^2} \frac{d}{dx}\left(e^{\alpha^2 x^2} u(x)\right)\right) + \lambda u(x) = 0, \quad x \in \mathbb{R}.$$

The corresponding eigenvalues are $\lambda_n = 2\alpha^2 n$. In contrast to the Hermite polynomials, the Hermite functions are well behaved with the decay property:

$$|\tilde{\mathrm{H}}_n(x)| \to 0, \quad \text{as } |x| \to \infty,$$

and the asymptotic formula with large n is

$$\tilde{\mathrm{H}}_n(x) \sim n^{-1/4} \cos\left(\sqrt{2n+1}x - \frac{n\pi}{2}\right).$$

Some sample graphs of the Hermite polynomials and the Hermite functions are presented in Figures 5.3 and 5.4, respectively.

We introduce the following inner product and the associated norm

$$(y, z)_{w_\alpha} = \int_{\mathbb{R}} y(x) z(x) w_\alpha(x) dx, \quad \|y\|_{w_\alpha} = (y, y)_{w_\alpha}^{1/2}, \qquad y, z \in L^2_{w_\alpha}(\mathbb{R}),$$

and also consider the weighted Sobolev space

$$H^r_{w_\alpha}(\mathbb{R}) = \left\{ y \,\middle|\, \frac{d^k y}{dx^k} \in L^2_{w_\alpha}(\mathbb{R}),\ 0 \le k \le r \right\},$$

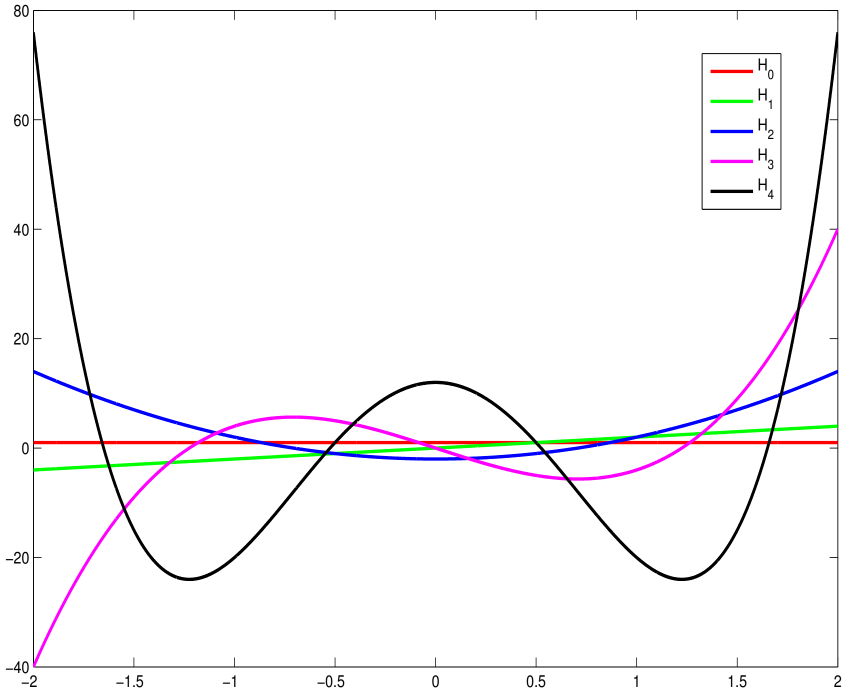


Figure 5.3: The first five Hermite polynomials $H_n(x)$, $n = 0, 1, 2, 3, 4$.

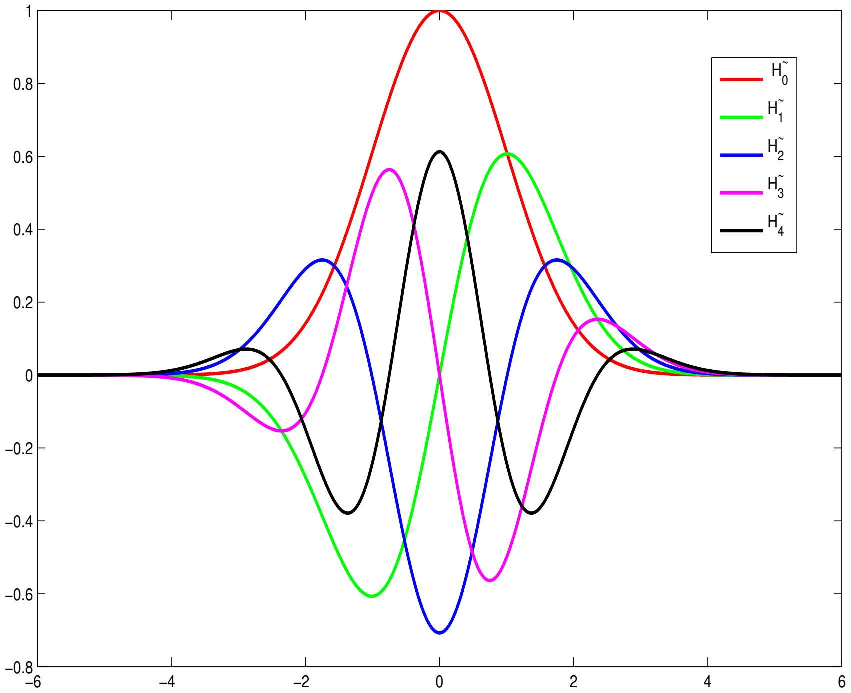


Figure 5.4: The first five Hermite functions $\tilde{\mathrm{H}}_n(x)$, $n = 0, 1, 2, 3, 4$.

equipped with the following semi-norm and norm, respectively,

$$|y|_{k,w_\alpha} = \|\frac{d^k y}{dx^k}\|_{w_\alpha}, \quad \|y\|_{r,w_\alpha} = \left(\sum_{k=0}^{r} |y|^2_{k,w_\alpha}\right)^{1/2}.$$

We note that the set of functions $\{\tilde{\mathrm{H}}_n(x), n \geq 0\}$ defines a $L^2_{w_\alpha}(\mathbb{R})$-orthogonal system with

$$(\tilde{\mathrm{H}}_n, \tilde{\mathrm{H}}_m)_{w_\alpha} = \frac{\sqrt{\pi}}{\alpha}\delta_{n,m}.$$

Therefore for all $y \in L^2_{w_\alpha}(\mathbb{R})$, we can write

$$y(x) = \sum_{n=0}^{\infty} \hat{y}_n \tilde{\mathrm{H}}_n(x),$$

with the coefficients

$$\hat{y}_n = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} y(x)\tilde{\mathrm{H}}_n(x) w_\alpha(x) dx, \quad n \geq 0.$$

We define

$$V_N = \{q(x) w_\alpha^{-1}(x) \,\big|\, q \in \mathbb{P}_N\},$$

and note that $V_N = \mathrm{span}\{\tilde{\mathrm{H}}_n(x),\ 0 \leq n \leq N\}$, where $\mathbb{P}_N$ is the set of polynomials of degree at most N. Therefore we can consider the $L^2_{w_\alpha}(\mathbb{R})$-orthogonal projection $P_N : L^2_{w_\alpha}(\mathbb{R}) \to V_N$, with

$$P_N y(x) = \sum_{n=0}^{N} \hat{y}_n \tilde{\mathrm{H}}_n(x).$$

In [45] the following theorem is proved, which is used in our work to estimate the approximation error in the space V_N.

Theorem 10. *For any $y \in H^r_{w_\alpha}(\mathbb{R})$ and $r \geq 0$,*

$$\|y - P_N y\|_{w_\alpha} \leq c(\alpha^2 N)^{-r/2} \|y\|_{r,w_\alpha},$$

where $c = (\frac{\alpha}{2^r\sqrt{\pi}})^{1/2}$.

This theorem also helps us to estimate the Hermite coefficients. We prove the following lemma.

Lemma 7. *For any $y \in H^r_{w_\alpha}(\mathbb{R})$, $r \geq 0$, and $n \geq 2$,*

$$|\hat{y}_n(t)| \leq \frac{\alpha^{1-r}}{\sqrt{\pi}} n^{-r/2} \|y(.,t)\|_{r,w_\alpha}.$$

Proof. Considering $n \geq 1$ and the orthogonality relation between Hermite functions, we can write the following inequality

$$\begin{aligned}
|\hat{y}_n(t)|^2 &\leq \sum_{k=n}^{\infty} |\hat{y}_k(t)|^2 = \frac{\alpha}{\sqrt{\pi}} \sum_{k=n}^{\infty} \frac{\sqrt{\pi}}{\alpha} |\hat{y}_k(t)|^2 = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} \left(\sum_{k=n}^{\infty} \hat{y}_k(t) \tilde{\mathrm{H}}_k(v) \right)^2 w_\alpha(v) dv \\
&= \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} \left(\sum_{k=0}^{\infty} \hat{y}_k(t) \tilde{\mathrm{H}}_k(v) - \sum_{k=0}^{n-1} \hat{y}_k(t) \tilde{\mathrm{H}}_k(v) \right)^2 w_\alpha(v) dv \\
&= \frac{\alpha}{\sqrt{\pi}} \|y - P_{n-1} y\|^2_{w_\alpha}.
\end{aligned}$$

By Theorem 10, we have

$$\|y - P_{n-1} y\|_{w_\alpha} \leq (\frac{\alpha}{2^r \sqrt{\pi}})^{1/2} (\alpha^2 (n-1))^{-r/2} \|y\|_{r, w_\alpha}.$$

Therefore,

$$|\hat{y}_n(t)| \leq \frac{2^{-r/2}}{\sqrt{\pi}} \alpha^{1-r} (n-1)^{-r/2} \|y(.,t)\|_{r, w_\alpha}.$$

Since for $n \geq 2$ we have $2(n-1) \geq n$, and consequently $2^{-r/2}(n-1)^{-r/2} \leq n^{-r/2}$, the following holds for $n \geq 2$,

$$|\hat{y}_n(t)| \leq \frac{\alpha^{1-r}}{\sqrt{\pi}} n^{-r/2} \|y(.,t)\|_{r, w_\alpha}.$$

□

To discretize the FP equation, we employ the following facts

$$\alpha x \tilde{\mathrm{H}}_n(x) = \sqrt{\frac{n+1}{2}}\, \tilde{\mathrm{H}}_{n+1}(x) + \sqrt{\frac{n}{2}}\, \tilde{\mathrm{H}}_{n-1}(x),$$

$$\frac{d}{dx} \tilde{\mathrm{H}}_n(x) = -\alpha \sqrt{2(n+1)}\, \tilde{\mathrm{H}}_{n+1}(x),$$

$$x \frac{d}{dx} \tilde{\mathrm{H}}_n(x) = -\sqrt{(n+1)(n+2)}\, \tilde{\mathrm{H}}_{n+2}(x) - (n+1) \tilde{\mathrm{H}}_n(x),$$

$$\frac{d^2}{dx^2} \tilde{\mathrm{H}}_n(x) = 2\alpha^2 \sqrt{(n+1)(n+2)}\, \tilde{\mathrm{H}}_{n+2}(x),$$

for $n \geq 0$, with $\tilde{\mathrm{H}}_j(x) = 0, j < 0$.

We also have

$$x H_n(x) = \frac{1}{2} H_{n+1}(x) + n H_{n-1}(x),$$

$$\frac{d}{dx} H_n(x) = 2n H_{n-1}(x),$$

$$x \frac{d}{dx} H_n(x) = n H_n(x) + 2n(n-1) H_{n-2}(x),$$

$$\frac{d^2}{dx^2}H_n(x) = 4n(n-1)H_{n-2}(x),$$

or equivalently,

$$\alpha x H_n(\alpha x) = \frac{1}{2}H_{n+1}(\alpha x) + nH_{n-1}(\alpha x),$$

$$\frac{d}{dx}H_n(\alpha x) = 2\alpha n H_{n-1}(\alpha x),$$

$$x\frac{d}{dx}H_n(\alpha x) = nH_n(\alpha x) + 2n(n-1)H_{n-2}(\alpha x),$$

$$\frac{d^2}{dx^2}H_n(\alpha x) = 4\alpha^2 n(n-1)H_{n-2}(\alpha x).$$

which provide the appropriate means to descretize the optimal control system.

We also prove the following lemma to discuss the conservativity of the discretized FP equation.

Lemma 8. *For $n \geq 1$*

$$\int_{\mathbb{R}} \tilde{H}_n(x)dx = 0.$$

Proof. Based on

$$\tilde{\mathrm{H}}_n(-x) = (-1)^n \tilde{\mathrm{H}}_n(x),$$

we see that $\tilde{\mathrm{H}}_n$ is an even function when n is even, and it is an odd function when n is odd. Therefore, it is clear that $\int_{\mathbb{R}} \tilde{\mathrm{H}}_n(x)dx = 0$ when n is odd. Assuming that n is even, and using the following fact

$$\int_0^x e^{-t^2}H_n(t)dt = H_{n-1}(0) - e^{-x^2}H_{n-1}(x),$$

we obtain

$$\begin{aligned}
\int_{\mathbb{R}} \tilde{\mathrm{H}}_n(x)dx &= \frac{1}{\sqrt{2^n n!}}\int_{\mathbb{R}} H_n(\alpha x)e^{-\alpha^2 x^2}dx \\
&= \frac{1}{\alpha\sqrt{2^n n!}}\int_{\mathbb{R}} H_n(t)e^{-t^2}dt \\
&= \frac{1}{\alpha\sqrt{2^n n!}}\left(\int_{-\infty}^0 H_n(t)e^{-t^2}dt + \int_0^\infty H_n(t)e^{-t^2}dt\right) \\
&= \frac{2}{\alpha\sqrt{2^n n!}}\lim_{x\to\infty}\int_0^x H_n(t)e^{-t^2}dt \\
&= \frac{2}{\alpha\sqrt{2^n n!}}\lim_{x\to\infty}\left(H_{n-1}(0) - e^{-x^2}H_{n-1}(x)\right).
\end{aligned}$$

Since H_{n-1} is an odd function, $H_{n-1}(0) = 0$ and the desired statement is proved. □

5.3 Discretization schemes

In this chapter, we investigate the Hermite spectral discretization of the FP equations, of both parabolic and hyperbolic types, defined on unbounded domains. The accuracy of the Hermite spectral method is proved by showing that the error decreases spectrally as the number of expansion terms increases. Furthermore, we investigate the conservativity of the solutions of the FP equations with Hermite discretization schemes. The accuracy of the discretization method is also investigated with numerical experiments.

5.3.1 The Fokker-Planck equation for Itō processes

We recall from Chapter 3 that in one dimension, the FP equation has the form

$$\partial_t f(x,t) = -\partial_x \left(b(x,t) f(x,t)\right) + \frac{1}{2}\partial_{xx} \left(a(x,t) f(x,t)\right), \tag{5.11}$$

for the Itō process given by the stochastic differential equation

$$dX(t) = b(X(t),t)dt + \sigma(X(t),t)\, dW(t)$$

with drift $b(X(t),t)$, dispersion $\sigma(X(t),t) = \sqrt{a(X(t),t)}$, and Wiener process $W(t)$. If the initial point $X(0)$ is a random variable which is distributed as $\rho(x)$, the FP model starts the evolution process from the initial probability density $f(x,0) = \rho(x)$.

We consider a FP model corresponding to a representative stochastic process given by the Ornstein-Uhlenbeck process, and for simplicity we focus on a case in which the function b is linear and a is constant. We have $b(x,t;u) = \gamma x + u$ and $a(x,t) = 2c$, where $\gamma < 0$, u and $c > 0$ are constants. In this case, the FP problem is given by

$$\partial_t f(x,t) = -\partial_x \left((\gamma x + u) f(x,t)\right) + c\partial_{xx} f(x,t), \quad \text{in } \mathbb{R} \times [0,T], \tag{5.12}$$

$$f(x,0) = \rho(x), \quad \text{in } \mathbb{R}. \tag{5.13}$$

As the first step of the Hermite discretization, the probability density f is approximated in the space of Hermite functions as follows

$$f(x,t) = \sum_{n=0}^{\infty} \hat{f}_n(t) \tilde{\mathrm{H}}_n(x).$$

The initial data $f(x,0) = \rho(x)$ is also represented in the Hermite functions space by $\rho(x) = \sum_{n=0}^{\infty} \hat{f}_n^0 \tilde{\mathrm{H}}_n(x)$, where

$$\hat{f}_n^0 = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} \rho(x) \tilde{\mathrm{H}}_n(x) w_\alpha(x) dx, \quad n \geq 0.$$

Introducing the Hermite expansion for f into the equation (5.12), for $n \geq 0$ we have

$$\frac{d}{dt}\hat{f}_n(t) = n\gamma \hat{f}_n(t) + \alpha u \sqrt{2n} \hat{f}_{n-1}(t) + (\gamma + 2\alpha^2 c)\sqrt{n(n-1)} \hat{f}_{n-2}(t), \tag{5.14}$$

with $\hat{f}_{-1} = 0$, $\hat{f}_{-2} = 0$.

The equation (5.14) represents an infinite system of ODEs. This system is truncated by considering the approximation

$$[\hat{f}_{\Delta,0}(t), \hat{f}_{\Delta,1}(t), \cdots, \hat{f}_{\Delta,N}(t)]$$

for

$$[\hat{f}_0(t), \hat{f}_1(t), \cdots].$$

Therefore, the system of ODEs which we solve is as follows

$$\begin{aligned} \tfrac{d}{dt}\hat{f}_{\Delta,n}(t) = n\gamma\hat{f}_{\Delta,n}(t) + \alpha\sqrt{2n}\hat{f}_{\Delta,n-1}(t) + (\gamma+2\alpha^2c)\sqrt{n(n-1)}\hat{f}_{\Delta,n-2}(t), \\ \hat{f}_{\Delta,n}(0) = \hat{\rho}_n, \end{aligned} \tag{5.15}$$

for $0 \leq n \leq N, 0 \leq t \leq T$, with $\hat{f}_{\Delta,i} = 0$, $i = -1, -2$. This corresponds to a Galerkin projection of $f(\cdot, t)$ onto the Hermite approximation space

$$V_N = \text{span}\{\tilde{\text{H}}_n(x),\ 0 \leq n \leq N\}.$$

The system (5.15) can be written in the following matrix form,

$$\frac{d\hat{f}_\Delta}{dt} = M_f\hat{f}_\Delta, \tag{5.16}$$

where

$$\hat{f}_\Delta = [\hat{f}_{\Delta,0}(t), \hat{f}_{\Delta,1}(t), \cdots, \hat{f}_{\Delta,N}(t)]^T,$$

and M_f is an $(N+1)\times(N+1)$ three-diagonal matrix with the elements

$$(M_f)_{ij} = \begin{cases} n\gamma, & i = j, \\ \alpha\sqrt{2n}, & i-j = 1, \\ (\gamma+2\alpha^2c)\sqrt{n(n-1)}, & i-j = 2, \\ 0, & \text{otherwise}, \end{cases} \qquad 1 \leq i,j \leq N+1,$$

where $n = i-1$. Notice that, the first row in M_f is zero.

5.3.1.1 Conservativity

Another important property of the numerical scheme is that the Hermite spectral discretization provides conservativeness. We prove the following

$$\int_{\mathbb{R}} f_\Delta(x,t)dx = \int_{\mathbb{R}} f_\Delta(x,0)dx, \quad t > 0.$$

First, we note that for any $t > 0$

$$\begin{aligned} \int_{\mathbb{R}} f_\Delta(x,t)dx &= \sum_{n=0}^{N} \hat{f}_{\Delta,n}(t) \int_{\mathbb{R}} \tilde{\text{H}}_n(x)\,dx \\ &= \hat{f}_{\Delta,0}(t) \int_{\mathbb{R}} \tilde{\text{H}}_0(x)\,dx. \end{aligned}$$

This is true because of Lemma 8 which states that $\int_{\mathbb{R}} \tilde{\mathrm{H}}_n(x)dx = 0$ for $n \geq 1$.

Noting that $\frac{d\hat{f}_\Delta}{dt} = M_f \hat{f}_\Delta$, and the fact that the first row of the matrix M_f is zero, we have

$$\hat{f}_{\Delta,0}(t) = \hat{f}_{\Delta,0}(0), \quad t > 0.$$

Therefore, we have

$$\begin{aligned}
\int_{\mathbb{R}} f_\Delta(x,t)dx &= \hat{f}_{\Delta,0}(t) \int_{\mathbb{R}} \tilde{\mathrm{H}}_0(x)\, dx \\
&= \hat{f}_{\Delta,0}(0) \int_{\mathbb{R}} \tilde{\mathrm{H}}_0(x)\, dx \\
&= \sum_{n=0}^{N} \hat{f}_{\Delta,n}(0) \int_{\mathbb{R}} \tilde{\mathrm{H}}_n(x)\, dx \\
&= \int_{\mathbb{R}} f_\Delta(x,0)\, dx.
\end{aligned}$$

5.3.1.2 Convergence analysis

Substituting the Hermite expansion into the FP equation results in an infinite system of linear ODEs. Corresponding to this system, there is a matrix M_∞, which is lower triangular. To have a practical scheme, we have to truncate this matrix, or equivalently, consider some truncated system of ODEs. However, this truncation is a source of error in our discretization scheme. Let $\|.\|_2$ be the Euclidean norm in $\mathbb{R}^{N+1}$. We have the following.

Lemma 9. *Assuming N is sufficiently large so that there is no error in the spectral representation of the initial data, and $f(\cdot,t) \in V_N$ for any $t \in [0,T]$, then*

$$\|\hat{f}_N - \hat{f}_\Delta\|_2 = 0.$$

That is, there will be no error for the truncation of the infinite ODE system.

Proof. No truncation error appears in calculating the Hermite coefficients $\hat{f}_n$ by solving the finite ODE system (5.16). This is because of the fact that the system (5.14) is uncoupled in the sense that for $m > n$ the value of $\hat{f}_n$ is independent of the value of $\hat{f}_m$. That is, $P_N f(\cdot,t) = f_\Delta(\cdot,t)$ for every $t \in [0,T]$. □

To analyze the accuracy of the scheme, we show that the approximation for the state variable is spectrally convergent.

Theorem 11. *If $f \in L^\infty(0,T; H^r_{w_\alpha}(\mathbb{R}))$, $r > 1$, then for all $t \in [0,T]$ the following holds*

$$\|f(\cdot,t) - f_\Delta(\cdot,t)\|^2_{w_\alpha} = O(N^{-r}).$$

Proof. We have

$$\begin{aligned} f(x,t) - f_\Delta(x,t) &= \sum_{n=0}^{\infty} \hat{f}_n(t)\tilde{\mathrm{H}}_n(x) - \sum_{n=0}^{N} \hat{f}_{\Delta,n}(t)\tilde{\mathrm{H}}_n(x) \\ &= \sum_{n=0}^{N} \left(\hat{f}_n(t) - \hat{f}_{\Delta,n}(t)\right)\tilde{\mathrm{H}}_n(x) + \sum_{n=N+1}^{\infty} \hat{f}_n(t)\tilde{\mathrm{H}}_n(x). \end{aligned}$$

From Lemma 9 we know that the first term in the last line of the equation above is zero, hence Lemma 7 gives us the following bound for the error

$$\begin{aligned} \|f(\cdot,t) - f_\Delta(\cdot,t)\|_{w_\alpha}^2 &= \int_{\mathbb{R}} \left[\sum_{n=N+1}^{\infty} \hat{f}_n(t)\tilde{\mathrm{H}}_n(x)\right]^2 w(x)dx \\ &= \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} |\hat{f}_n(t)|^2 \\ &\leq \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} \frac{\alpha^{2-2r}}{\pi} n^{-r} \|f(\cdot,t)\|_{r,w_\alpha}^2. \end{aligned}$$

Therefore, we have $\|f(\cdot,t) - f_\Delta(\cdot,t)\|_{w_\alpha}^2 = O(N^{-r})$. □

The following lemma provides an appropriate means to show that the Hermite discretization method is stable.

Lemma 10. *Let $\hat{y}(t)$ be the solution to*

$$\frac{d}{dt}\hat{y}(t) = M_f \hat{y}, \qquad \hat{y}(0) = \hat{y}_0.$$

Then there exists a constant C_N such that for all $t > 0$

$$\|\hat{y}(t)\|_2 \leq C_N \|\hat{y}_0\|_2.$$

Proof. Since the matrix M_f is triangular, it has $N+1$ distinct eigenvalues $\lambda_n = n\gamma$, $n = 0, 1, \cdots, N$, which are the diagonal elements of M_f. Therefore, M_f is diagonalizable and can be decomposed as $M_f = S^{-1}DS$, where $D = diag(\lambda_n)_{n=0}^N$. Hence, the system of ODEs has the solution

$$\hat{y}(t) = e^{M_f t}\hat{y}_0,$$

which implies the following

$$\|\hat{y}(t)\|_2 \leq \|S^{-1}\|_2 \|e^{Dt}\|_2 \|S\|_2 \|\hat{y}_0\|_2.$$

Since $\gamma < 0$, we have $e^{2\lambda_n t} \leq 1$, $n = 0, 1, \cdots, N$, and consequently

$$\|e^{Dt}\|_2 = \sigma_{\max}(e^{Dt}) = \sqrt{\lambda_{\max}(e^{2Dt})} \leq 1.$$

It is easy to show that the matrices S and S^{-1} are also lower triangular. Since S is consist of the eigenvectors of M_f, it can be constructed in such a way that all diagonal elements are 1. Defining $\tilde{s} := \|S\|_{\max}$ we have

$$\|S\|_2 \leq (N+1)\|S\|_{\max} = (N+1)\tilde{s}.$$

Furthermore, in [69] it is proved that

$$\|S^{-1}\|_\infty \leq (\tilde{s}+1)^N,$$

which results in

$$\|S^{-1}\|_2 \leq \sqrt{N+1}\|S^{-1}\|_\infty = \sqrt{N+1}(\tilde{s}+1)^N.$$

Therefore, we have

$$\|\hat{y}(t)\|_2 \leq C_N\|\hat{y}_0\|_2.$$

where $C_N = (N+1)^{3/2}(\tilde{s}+1)^{N+1}$.

□

Based on Lemma 10, we have the following stability result.

Theorem 12. *There exists a constant C_N such that for all $t > 0$*

$$\|f_\Delta(.,t)\|_{w_\alpha} \leq C_N\|\hat{f}_0\|_2.$$

Proof. We have

$$\begin{aligned}\|f_\Delta(.,t)\|^2_{w_\alpha} &= \int_{\mathbb{R}} (f_\Delta)^2 w_\alpha(x)dx = \int_{\mathbb{R}} \left(\sum_{n=0}^{N} \hat{f}_{\Delta,n}(t)\tilde{\mathrm{H}}_n(x)\right)^2 w_\alpha(x)dx \\ &= \frac{\sqrt{\pi}}{\alpha}\sum_{n=0}^{N}(\hat{f}_{\Delta,n}(t))^2 = \frac{\sqrt{\pi}}{\alpha}\|\hat{f}_\Delta(t)\|_2^2 \leq C\|\hat{f}_0\|_2^2.\end{aligned}$$

□

5.3.1.3 Numerical experiments

Since the system of ODEs which we need to solve in order to obtain the numerical solutions are first order linear systems, there exists no time discretization in our numerical scheme. That is, we can calculate the Hermite expansion coefficients analytically and without any time discretization error. The triangular structure of the matrices of coefficients with distinct eigenvalues, makes it possible to decompose the mentioned matrices and solve the system of ODEs simply by matrix products. Therefore, the errors presented in this section are only induced by spatial discretization.

In [45] it is stated that the Hermite spectral method does not provide good resolution for all scaling factor α. It is thence proved that to approximate Gaussian type functions e^{-sx^2}, the scaling factor α must satisfy $0 < \alpha < \sqrt{2s}$. Consider the forward FP equation

$$\partial_t f(x,t) - c\partial_{xx} f(x,t) + \partial_x \left((\gamma x + u) f(x,t)\right) = 0.$$

It is easy to see that the stationary solution, which satisfies $\partial_t f = 0$, or equivalently

$$\partial_x \left(c\partial_x f - (\gamma x + u) f\right) = 0,$$

is as follows

$$f(x) = C_0 \exp(\frac{\gamma}{2c}x^2 + \frac{u}{c}x),$$

where C_0 is a constant. Comparing the stationary solution with the weight function $w_\alpha(x)$, while the control variable $u = 0$, motivates us to set $\alpha = \sqrt{\frac{-\gamma}{2c}}$. This choice satisfies the condition mentioned in [45], and seems to be the best option since to find the optimal scaling factor is still an open problem.

To illustrate the importance of choosing a proper scaling factor, consider Case 1 with a known exact solution for the following FP equation

$$\partial_t f - \partial_{xx} f - \partial_x (xf) = 0,$$

with the initial condition

$$f(x,0) = e^{(-\frac{x^2}{2})} \left(1 + \cos(\frac{\pi}{2}x) \exp(\frac{\pi^2}{8})\right).$$

The exact solution of this problem is given by

$$f(x,t) = e^{(-\frac{x^2}{2})} \left(1 + \cos(\frac{\pi}{2}xe^{-t}) \exp(\frac{\pi^2}{8})e^{-2t}\right).$$

Since the parameters of the FP equation are $c = 1$, $\gamma = -1$, and $u = 0$, we set $\alpha := \sqrt{\frac{-\gamma}{2c}} = \frac{1}{\sqrt{2}} \approx 0.7071$. Figure 5.5 illustrates how different values for the scaling factor may lead to different approximations for a given N. However, as mentioned in [45] the Hermite approximation is accurate in solving for the asymptotic solution also without an optimal α.

In Table 5.4, we see how fast the error decreases when t increases. After reaching to the equilibrium solution the error remains at the value of the machine error. We can investigate more about Hermite discretization with this experiment. Table 5.2 shows the decay of the error regarding increasing N.

Since in our Hermite spectral discretization, the initial condition of the differential equation has to be mapped into the approximation space V_N, if N is not large enough to have a precise representation of the initial data, one cannot expect a satisfactory numerical result. However, in Figure 5.6 we see that for the problems dealing with a Gaussian type function, the influence of the error in representing the initial data becomes negligible along time evolution.

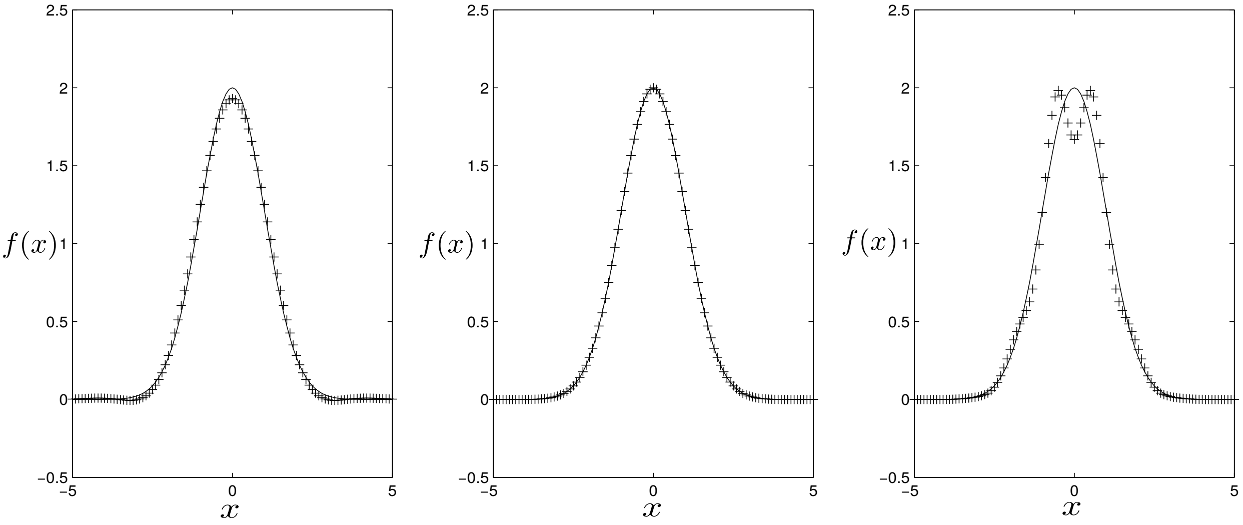


Figure 5.5: Case 1: numerical (cross-marks) and exact solution (solid line) to the FP equation with different scaling factors; left: $\alpha = 0.4$, middle: $\alpha = 0.7071$, right: $\alpha = 1$; $N = 10$ and $T = 10$.

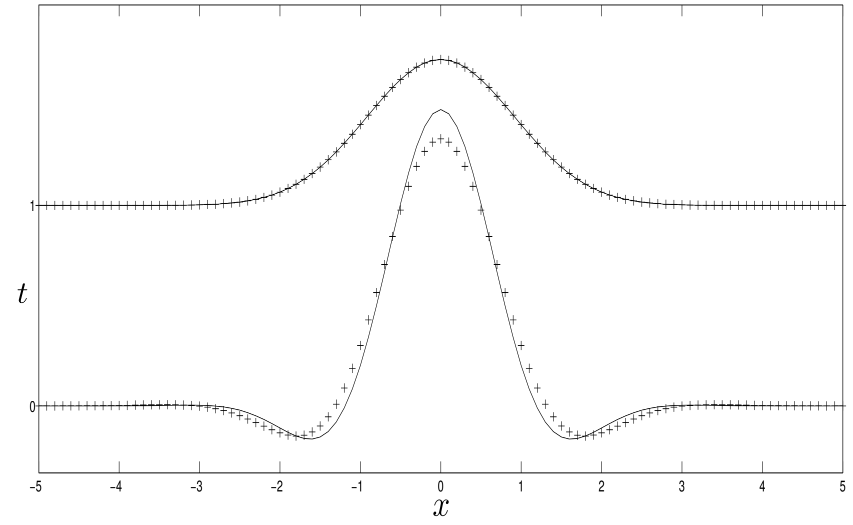
T	$\|f_\Delta - f_{exact}\|_{L^2}$
1	2.0101e-11
2	1.2132e-16
3	3.0412e-18
4	3.0428e-18
5	3.0450e-18

Table 5.1: Case 1: decay of the solution error at final time when T increases; $N = 10$, $\alpha = 0.7071$.

N	$\|f_\Delta - f_{exact}\|_{L^2}$
5	4.2193e-07
10	2.0101e-11
15	1.0275e-14
20	2.9166e-18

Table 5.2: Decay of the error in case 1 when N increases; $T = 1$, $\alpha = 0.7071$

To examine the Hermite discretization scheme concerning positivity preserving and conservativity, we introduce Case 2. In this experiment, we can also compare the approximated solution with the exact solution of an FP equation with a non-zero u, which is presented in [6] and will be emphasized in Chapter 7 where u plays the role



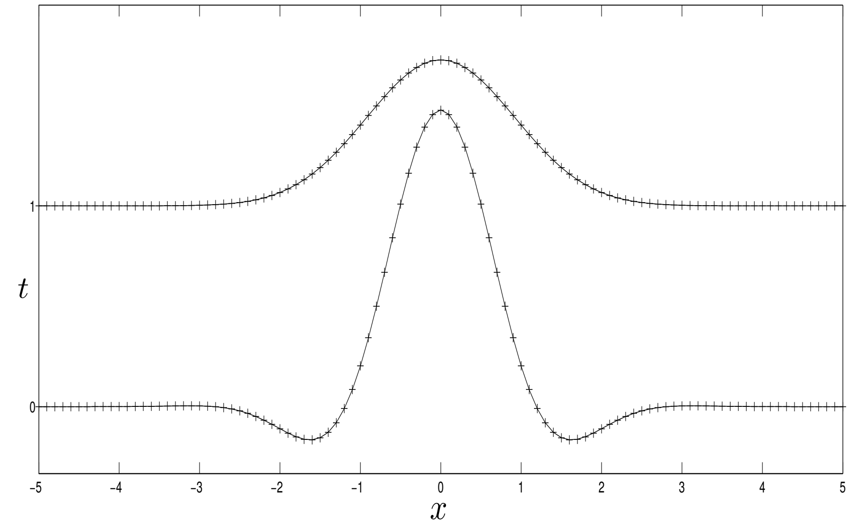


Figure 5.6: Case 1: Accurate approximation for the solution at $T = 1$ (top graph), even if the initial solution is not well approximated (bottom graph). Top figure: $N = 5$, bottom figure: $N = 10$. Cross-marks represent the numerical solution and the solid lines represent the exact solution.

of a control variable. The exact solution of the FP equation

$$\partial_t f(x,t) - c\partial_{xx} f(x,t) + \partial_x \left((\gamma x + u) f(x,t)\right) = 0$$

with the initial condition

$$f_0(x) = \delta(x)$$

is a Gaussian distribution with mean $\mu(t,u) = -u/\gamma + (u/\gamma)e^{\gamma t}$ and variance $\bar{\sigma}^2(t) = -c/\gamma(1 - e^{2\gamma t})$, that is

$$f(x,t,u) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2(t)}} \exp\left(-\frac{(x-\mu(t,u))^2}{2\bar{\sigma}^2(t)}\right).$$

Since it is impossible to represent the Dirac delta function $\delta(0)$ by Hermite functions, we apply a temporal shift in the exact solution in order to have a Gaussian function as the initial condition. In Figure 5.7, time $t = 1$ has been considered to be the starting time of the process which evolves under the action of the control $u = 2$. We observe how fast the approximation becomes accurate as N increases.

Since in this case, $u \neq 0$ and consequently the stationary solution is not centered at zero, it becomes harder to deal with a proper choice of the scaling factor α. By trying different values of α, we gain the best estimate corresponding to $\alpha = 0.7$. however, the error estimate presented in Table 5.3 is not as perfect as the estimation in Case 1, which is due to considering a constant scaling factor instead of a time dependent one. The idea of the time dependent scaling factor is discussed in [75], while one can also think about inserting a translating factor into the Hermite functions to treat the non zero-centered Gaussian functions. This strategy is applied in [74].

Further, Table 5.3 verifies that when the number of the expansion terms is large enough to have a non-negative representation of the initial PDF, the discretization scheme leads to a non-negative solution of the forward FP equation. We observe that the property $\int_{\mathbb{R}} f(x,t) = 1$, $t \geq 0$, is perfectly preserved independent of the number of expansion terms.

N	Error	$\min_{x\in\mathbb{R}} f_\Delta(x,0)$	$\min_{x\in\mathbb{R}} f_\Delta(x,5)$	$\int_{\mathbb{R}} f_\Delta(x,0)dx$	$\int_{\mathbb{R}} f_\Delta(x,5)dx$
5	0.4307	-3.1391e-04	-0.2628	1.0000	1.0000
10	0.0403	-5.8202e-05	-0.0095	1.0000	1.0000
15	0.0054	-2.3574e-06	-8.5076e-05	1.0000	1.0000
20	0.0053	-2.7610e-08	-2.5495e-07	1.0000	1.0000
25	0.0053	-7.8068e-10	-1.2459e-11	1.0000	1.0000
30	0.0053	-2.9185e-11	0	1.0000	1.0000

Table 5.3: Case 2: Positivity preserving and conservativity of Hermite discretization.

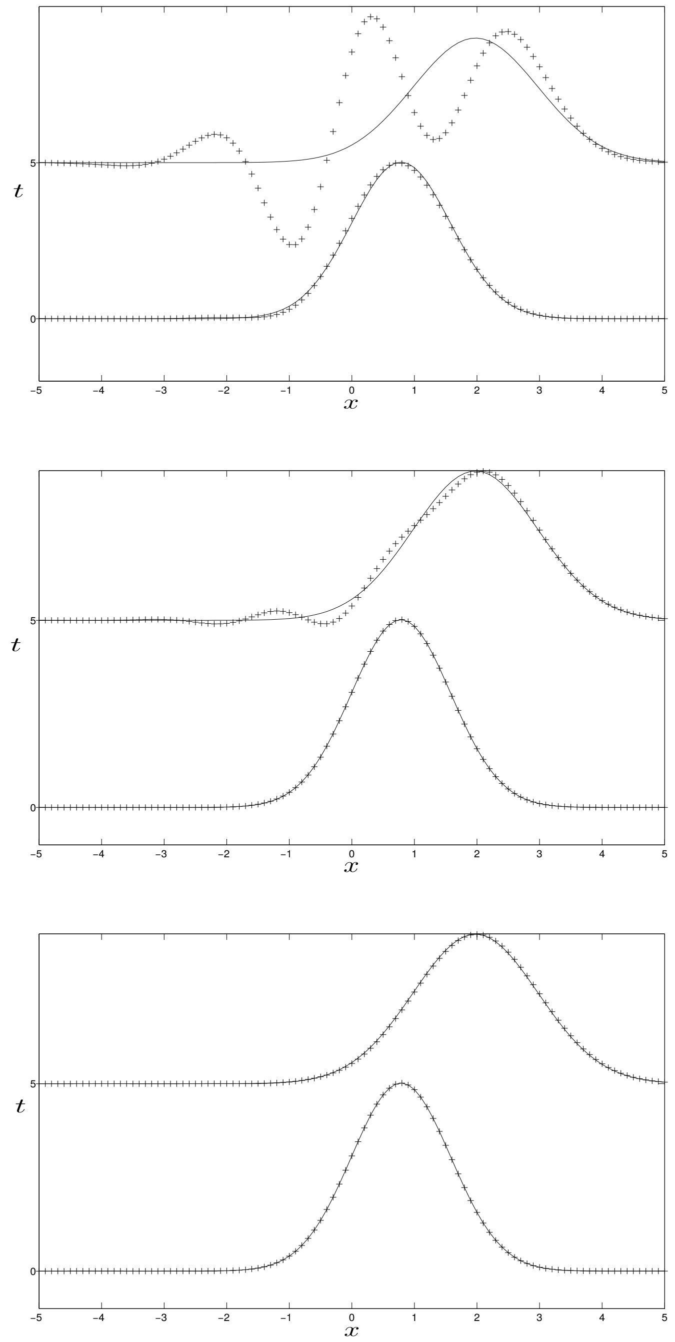


Figure 5.7: Case 2: Approximations of the initial PDF and the solution at $T = 5$. Top figure: $N = 5$, middle figure: $N = 10$, bottom figure: $N = 15$. Cross-marks represent the numerical solution and the solid lines represent the exact solution.

5.3.2 The Fokker-Planck equation for PDP processes

To apply Hermite spectral discretization, we consider another class of FP equations, where the PDF corresponds to a piecewise deterministic process. We consider the PDP model introduced in Chapter 3, that is a first-order system of ordinary differential equations, where the driving dynamics-function is chosen by a renewal process.

We recall from Chapter 3 that the time evolution of the PDFs of the piecewise deterministic process

$$\frac{d}{dt}X(t) = A_{\mathscr{S}(t)}\left(X(t)\right), \quad t \in [t_0, \infty), \tag{5.17}$$

is governed by the following FP hyperbolic system,

$$\partial_t f_s(x,t) + \partial_x(A_s(x) f_s(x,t)) = \sum_{j=1}^{S} \mathcal{Q}_{sj} f_j(x,t), \quad s \in \mathbb{S}, \tag{5.18}$$

where $\mathcal{Q}_{sj}$, $s, j \in \mathbb{S}$, are the components of the transition matrix $\mathcal{Q}$. The initial conditions for the PDFs of the FP system are given as follows

$$f_s(x,0) = f_s^0(x), \quad s \in \mathbb{S}, \tag{5.19}$$

where $f_s^0(x) \geq 0$, $x \in \mathbb{R}^d$, $\sum_{s=1}^{S} \int_{\mathbb{R}}^{d} f_s^0(x)\,dx = 1$.

For Hermite spectral discretization, we focus on the following structure: $A_s(x) = a_s x + c_s$, where a_s and c_s are constants $s \in \mathbb{S}$. This choice of the dynamics A_s includes the dissipative process subject to dichotomic noise considered in [8]. In this case, the proposed scheme can be validated by comparison with some exact solutions.

After replacing A_s, equation (5.18) becomes

$$\partial_t f_s = -a_s x\, \partial_x f_s - c_s \partial_x f_s + (\mathcal{Q}_{ss} - a_s) f_s + \sum_{\substack{j=1 \\ j \neq s}}^{S} \mathcal{Q}_{sj} f_j, \quad s \in \mathbb{S}. \tag{5.20}$$

We denote the approximation of $f_s(\cdot,t)$ with $f_\Delta^s(\cdot,t) \in V_N$, where

$$f_\Delta^s(x,t) = \sum_{n=0}^{N} \hat{f}_n^s(t) \tilde{\mathrm{H}}_n(x). \tag{5.21}$$

We insert (5.21) into (5.20) and obtain

$$\frac{d}{dt}\hat{f}_n^s(t) = d_n^s \hat{f}_n^s(t) + l_n^s \hat{f}_{n-1}^s(t) + h_n^s \hat{f}_{n-2}^s(t) + \sum_{\substack{j=1 \\ j \neq s}}^{S} k_j^s \hat{f}_n^j(t), \quad s \in \mathbb{S}, \tag{5.22}$$

where $d_n^s = \mathcal{Q}_{ss} + a_s n$, $l_n^s = \alpha\sqrt{2n}(c_s)$, $h_n^s = a_s\sqrt{n(n-1)}$, $k_j^s = \mathcal{Q}_{sj}$ and $\hat{f}_{-1}^s = \hat{f}_{-2}^s = 0$. Equation (5.22) is valid for $0 \leq n \leq N$ and gives the following system of ODEs

$$\frac{d\hat{f}_\Delta(t)}{dt} = M_f(t)\hat{f}_\Delta(t), \tag{5.23}$$

where

$$\hat{f}_\Delta = [\hat{f}_0^1, \cdots, \hat{f}_N^1, \hat{f}_0^2, \cdots, \hat{f}_N^2, \cdots, \hat{f}_0^S, \cdots, \hat{f}_N^S]^T.$$

The matrix M_f is a $S(N+1) \times S(N+1)$ sparse matrix as follows

$$M_f = \begin{bmatrix}
d_0^1 & & & & & k_2^1 & & & & & & k_S^1 & & & & \\
l_1^1 & d_1^1 & & & & & k_2^1 & & & & & & k_S^1 & & & \\
h_2^1 & l_2^1 & d_2^1 & & & & & k_2^1 & & & \cdots & & & k_S^1 & & \\
& \ddots & \ddots & \ddots & & & & & \ddots & & & & & & \ddots & \\
& & h_N^1 & l_N^1 & d_N^1 & & & & & k_2^1 & & & & & & k_S^1 \\
k_1^2 & & & & & d_0^2 & & & & & & k_S^2 & & & & \\
& k_1^2 & & & & l_1^2 & d_1^2 & & & & & & k_S^2 & & & \\
& & k_1^2 & & & h_2^2 & l_2^2 & d_2^2 & & & \cdots & & & k_S^2 & & \\
& & & \ddots & & & \ddots & \ddots & \ddots & & & & & & \ddots & \\
& & & & k_1^2 & & & h_N^2 & l_N^2 & d_N^2 & & & & & & k_S^2 \\
& & \vdots & & & & & \vdots & & & \ddots & & & \vdots & & \\
k_1^S & & & & & k_2^S & & & & & & d_0^S & & & & \\
& k_1^S & & & & & k_2^S & & & & & l_1^S & d_1^S & & & \\
& & k_1^S & & & & & k_2^S & & & \cdots & h_2^S & l_2^S & d_2^S & & \\
& & & \ddots & & & & & \ddots & & & & \ddots & \ddots & \ddots & \\
& & & & k_1^S & & & & & k_2^S & & & & h_N^S & l_N^S & d_N^S
\end{bmatrix}.$$

The initial data $f_s(x,0) = f_s^0(x)$, $s \in \mathbb{S}$, needs also to be represented in the Hermite functions space by the following

$$\bar{f}_\Delta^s(x) = \sum_{n=0}^{N} \bar{f}_n^s \tilde{\mathrm{H}}_n(x),$$

where

$$\bar{f}_n^s = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} f_s^0(x) \tilde{\mathrm{H}}_n(x) w_\alpha(x) dx, \quad n \geq 0.$$

Let

$$\bar{f}_\Delta = [\bar{f}_0^1, \cdots, \bar{f}_N^1, \bar{f}_0^2, \cdots, \bar{f}_N^2, \cdots, \bar{f}_0^S, \cdots, \bar{f}_N^S]^T,$$

Now the PDE problem of approximating the solution of the forward FP equations reduces to the problem of solving the following linear first-order system of ODEs,

$$\frac{d\hat{f}_\Delta(t)}{dt} = M_f(t)\hat{f}_\Delta(t), \quad \hat{f}_\Delta(0) = \bar{f}_\Delta.$$

We notice that the solution $\hat{f}_\Delta$ is given by

$$\hat{f}_\Delta(t) = \exp(M_f t)\, \bar{f}_\Delta.$$

5.3.2.1 Conservativity

Here, we deal with conservativity of the discretization scheme, which is another essential property for the discretized FP system. In fact, we require

$$\sum_{s=1}^{S}\int_{\mathbb{R}} f_{\Delta}^{s}(x,t)\,dx = \sum_{s=1}^{S}\int_{\mathbb{R}} f_{s}^{0}(x)\,dx, \quad t>0.$$

Lemma 8 makes it possible to prove the following theorem.

Theorem 13. *The Hermite spectral discretization of the FP system is conservative, that is,*

$$\frac{d}{dt}\left(\sum_{s=1}^{S}\int_{\mathbb{R}} f_{\Delta}^{s}(x,t)\,dx\right) = 0, \quad t>0.$$

Proof. We have

$$\begin{aligned}
\frac{d}{dt}\left(\sum_{s=1}^{S}\int_{\mathbb{R}} f_{\Delta}^{s}(x,t)\,dx\right) &= \frac{d}{dt}\sum_{s=1}^{S}\int_{\mathbb{R}}\left(\sum_{n=0}^{N}\hat{f}_{n}^{s}(t)\tilde{\mathrm{H}}_{n}(x)\right)dx \\
&= \sum_{s=1}^{S}\sum_{n=0}^{N}\frac{d}{dt}\hat{f}_{n}^{s}(t)\int_{\mathbb{R}}\tilde{\mathrm{H}}_{n}(x)dx \\
&= \sum_{s=1}^{S}\frac{d}{dt}\hat{f}_{0}^{s}(t)\int_{\mathbb{R}}\tilde{\mathrm{H}}_{0}(x)dx \\
&= \int_{\mathbb{R}}\tilde{\mathrm{H}}_{0}(x)dx\sum_{s=1}^{S}\frac{d}{dt}\hat{f}_{0}^{s}(t).
\end{aligned}$$

Notice that, we have used Lemma 8, which says that the integral of all Hermite functions over $\mathbb{R}$ is zero except $\tilde{\mathrm{H}}_0$. Now, we have the following

$$\begin{aligned}
\sum_{s=1}^{S}\frac{d}{dt}\hat{f}_{0}^{s}(t) &= \sum_{s=1}^{S}\left(\mathcal{Q}_{ss}\hat{f}_{0}^{s}(t)+\sum_{\substack{j=1\\ j\neq s}}^{S}\mathcal{Q}_{sj}\hat{f}_{0}^{j}(t)\right) \\
&= \sum_{j=1}^{S}\sum_{s=1}^{S}\mathcal{Q}_{sj}\hat{f}_{0}^{j}(t) \\
&= \sum_{j=1}^{S}\hat{f}_{0}^{j}(t)\sum_{s=1}^{S}\mathcal{Q}_{sj} \\
&= \sum_{j=1}^{S}\hat{f}_{0}^{j}(t)\times 0 \\
&= 0.
\end{aligned}$$

□

This proves conservativeness of the FP system. In particular, we have

$$\sum_{s=1}^{S} \int_{\mathbb{R}} f_{\Delta}^{s}(x,t)\, dx = \sum_{s=1}^{S} \int_{\mathbb{R}} f_{s}^{0}(x)\, dx = 1, \quad t > 0.$$

5.3.2.2 Convergence analysis

In this section, we discuss the approximation properties of our Hermite discretization scheme applied to the FP optimality system (6.12). We show that the approximations for the state variable is spectrally convergent.

Theorem 14. *If $f_s \in L^{\infty}(0,T; H^{r}_{w_\alpha}(\mathbb{R}))$ for $s \in \mathbb{S}$ and $r > 1$, then for all $t \in [0,T]$ the following holds*

$$\|f_s(\cdot,t) - f_{\Delta}^{s}(\cdot,t)\|_{w_\alpha}^{2} = O(N^{-r}).$$

Proof. For any $s \in \mathbb{S}$, we have

$$\begin{aligned} f_s(x,t) - f_{\Delta}^{s}(x,t) &= \sum_{n=0}^{\infty} \hat{f}_{n}^{s}(t)\tilde{\mathrm{H}}_n(x) - \sum_{n=0}^{N} \hat{f}_{n}^{s}(t)\tilde{\mathrm{H}}_n(x) \\ &= \sum_{n=N+1}^{\infty} \hat{f}_{n}^{s}(t)\tilde{\mathrm{H}}_n(x). \end{aligned}$$

Notice that for constant controls no truncation error appears in calculating the Hermite coefficients $\hat{f}_{n}^{s}$ by solving the ODE system (5.16). This is because of the fact that for $m > n$ the value of $\hat{f}_{n}^{s}$ is independent of the value of $\hat{f}_{m}^{s}$. That is, $P_N f_s(\cdot,t) = f_{\Delta}^{s}(\cdot,t)$ for every $t \in [0,T]$. So we have the following bound for the error

$$\begin{aligned} \|f_s(\cdot,t) - f_{\Delta}^{s}(\cdot,t)\|_{w_\alpha}^{2} &= \int_{\mathbb{R}} \left[\sum_{n=N+1}^{\infty} \hat{f}_{n}^{s}(t)\tilde{\mathrm{H}}_n(x) \right]^2 w(x)dx \\ &= \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} |\hat{f}_{n}^{s}(t)|^2 \\ &\leq \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} \frac{\alpha^{2-2r}}{\pi} n^{-r} \|f_s(\cdot,t)\|_{r,w_\alpha}^{2}. \end{aligned}$$

Therefore, we have $\|f_s(\cdot,t) - f_{\Delta}^{s}(\cdot,t)\|_{w_\alpha}^{2} = O(N^{-r})$.

□

5.3.2.3 Numerical experiments

The forward equations are expected to approach the steady equilibrium state asymptotically. As discussed in [4] and [82], the steady state is known in the case of a PDP with dichotomic noise where the dynamics corresponds to a linear filter. We have $A_1(x) = -a_1 x + c_1$ and $A_2(x) = -a_2 x + c_2$ and the following FP model

$$\partial_t f_1(x,t) + \partial_x\left((-a_1 x + c_1) f_1(x,t)\right) = \mathcal{Q}_{11} f_1(x,t) + \mathcal{Q}_{12} f_2(x,t), \tag{5.24}$$

$$\partial_t f_2(x,t) + \partial_x\left((-a_2 x + c_2) f_2(x,t)\right) = \mathcal{Q}_{21} f_1(x,t) + \mathcal{Q}_{22} f_2(x,t). \tag{5.25}$$

In particular, we choose $a_1 = a_2 = \gamma$, $c_1 = W$, $c_2 = -W$, $\mathcal{Q}_{11} = \mathcal{Q}_{22} = -\mu$ and $\mathcal{Q}_{12} = \mathcal{Q}_{21} = \mu$.

Defining $\delta = \mu/\gamma$, one has $\lim_{t\to\infty}(f_1 + f_2) = P_{eq}$ where P_{eq} is the equilibrium density given by

$$P_{eq}(x) = \frac{\gamma}{W\sqrt{\pi}} \frac{\Gamma(\delta + 1/2)}{\Gamma(\delta)} \left(1 - (\frac{x\gamma}{W})^2\right)^{\delta - 1}.$$

The equilibrium density is defined in the interval $\Omega = [-W/\gamma, W/\gamma]$, and is zero outside of Ω. $P_{eq}(x)$ has the transition point at $\delta = 1$. The point $x = 0$ is the maximum if $\delta > 1$, and is the minimum if $\delta < 1$. We consider these three cases regarding the value of δ, and compare with P_{eq} at time $t = 20$. Let $W = 1$, and consider the initial conditions

$$f_1^0(x) = f_2^0(x) = \frac{1}{\sqrt{8\pi\sigma_0^2}} \exp(\frac{-x^2}{2\sigma_0^2}),$$

with $\sigma_0^2 = 0.5$. Notice that these initial condition functions are infinitely differentiable.

In the first case, Case 1, we set $\mu = 2$ and $\gamma = 0.25$, so that $\delta = 8 > 1$ and $\Omega = [-4, 4]$. Figure 5.8 along with Table 5.4 show the spectral rate of convergence for $\alpha = 0.7$. We obtain evidence of accurate approximation of Hermite spectral discretization regarding the Gaussian-type functions. Therefore, for the setting with $\delta \gg 1$ just a few expansion terms are adequate to have an accurate approximation, since the equilibrium solution presents a Gaussian shape in this case [5].

N	$\Vert f_\Delta^1 + f_\Delta^2 - P_{eq}\Vert_{L^2}$	$\int_{\mathbb{R}} (f_\Delta^1 + f_\Delta^2)(x,0)\,dx$	$\int_{\mathbb{R}} (f_\Delta^1 + f_\Delta^2)(x,30)\,dx$
3	2.04e-2	1.0000	1.0000
10	1.36e-3	1.0000	1.0000
15	4.90e-4	1.0000	1.0000
30	2.22e-5	1.0000	1.0000
60	7.34e-6	1.0000	1.0000
120	5.86e-7	1.0000	1.0000

Table 5.4: Conservativity and spectral convergence in Case 1; $\alpha = 0.7$.

In the second case, Case 2, we set $\mu = 0.2$ and $\gamma = 1$ to have $\delta = 0.2 < 1$. The domain of definition of P_{eq} is $\Omega = [-1, 1]$. Since we have singularity points at the boundary of Ω, we cannot expect to observe a good approximation in the space of continuous Hermite functions when N is not very large. In this case, the scaling

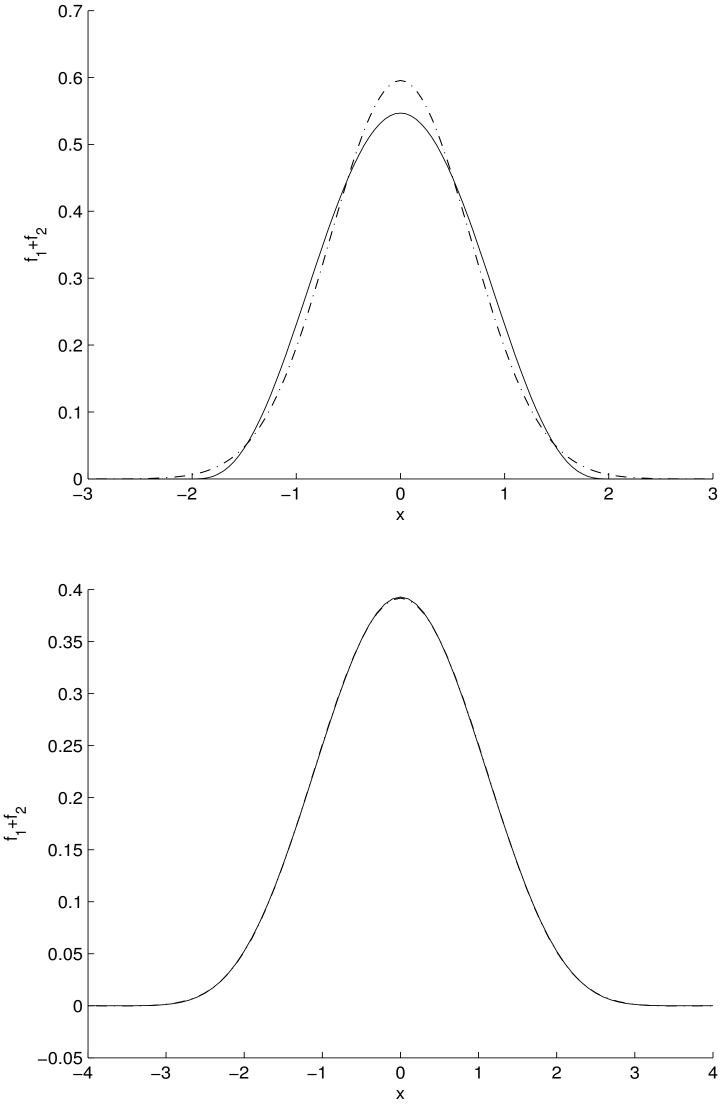


Figure 5.8: Case 1, $\delta = 8$; Numerical (dash-dot line) and equilibrium solution (solid line) to the summation of states $f_1 + f_2$; top: $N = 3$, bottom: $N = 10$; $\alpha = 0.7$.

factor plays an essential role, which we notice by plotting the solutions corresponding to different values of α. For instance, in Figure 5.9 we plot the results for $\alpha = 0.7$. Figure 5.10 shows that with a proper scaling factor, we can improve with Hermite functions to represent singularities. However, the number of expansion terms must be large in this case to remove undesired oscillations which are due to singularities.

In the third case, Case 3, we choose $\mu = \gamma = 0.5$, and consequently $\delta = 1$ and $\Omega = [-2, 2]$. We again encounter a discontinuous solution, which is $P_{eq}(x) = 0.25$ for $x \in \Omega$, and $P_{eq}(x) = 0$ for $x \notin \Omega$. In Figure 5.11, we can see that we need a high resolution to overcome this discontinuity. Figure 5.12 illustrates the effect of the scaling factor in improving the accuracy for a fixed number of expansion terms.

The results show that the Hermite approximation attempts to capture the singularities, which are at the boundary points of the domain of definition in the cases where $\delta < 1$ and $\delta = 1$. It leads to an oscillation near the boundary points, and results in negative values. To illustrate the influence of non-differentiability on the numerical results, we present the solutions for Case 2 at early times $t = 1$ and $t = 2$. It can be seen that undesired oscillations appear near the boundary of the domain of definition as time proceeds. The same behavior is expected for Case 3, which is confirmed by the solutions plotted at times $t = 2$ and $t = 3$ in Figure 5.14. Notice that in these cases, we reach the steady state at time $t \approx 10$. In spite of non-negativity, it can be observed and numerically verified that in all cases the discretization scheme is conservative. That is, the integral of the summation of the PDFs over the whole domain is preserved as time proceeds.

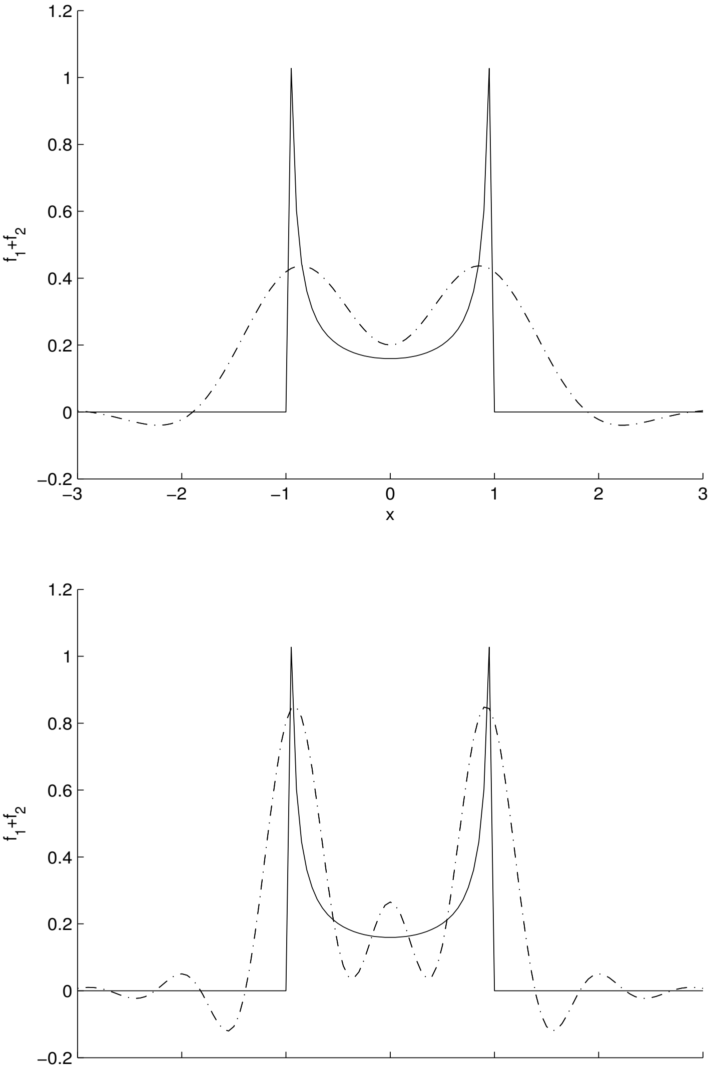


Figure 5.9: Case 2, $\delta < 1$; Numerical (dash-dot line) and equilibrium solution (solid line) to the summation of states $f_1 + f_2$; top: $N = 10$, bottom: $N = 50$; $\alpha = 0.7$.

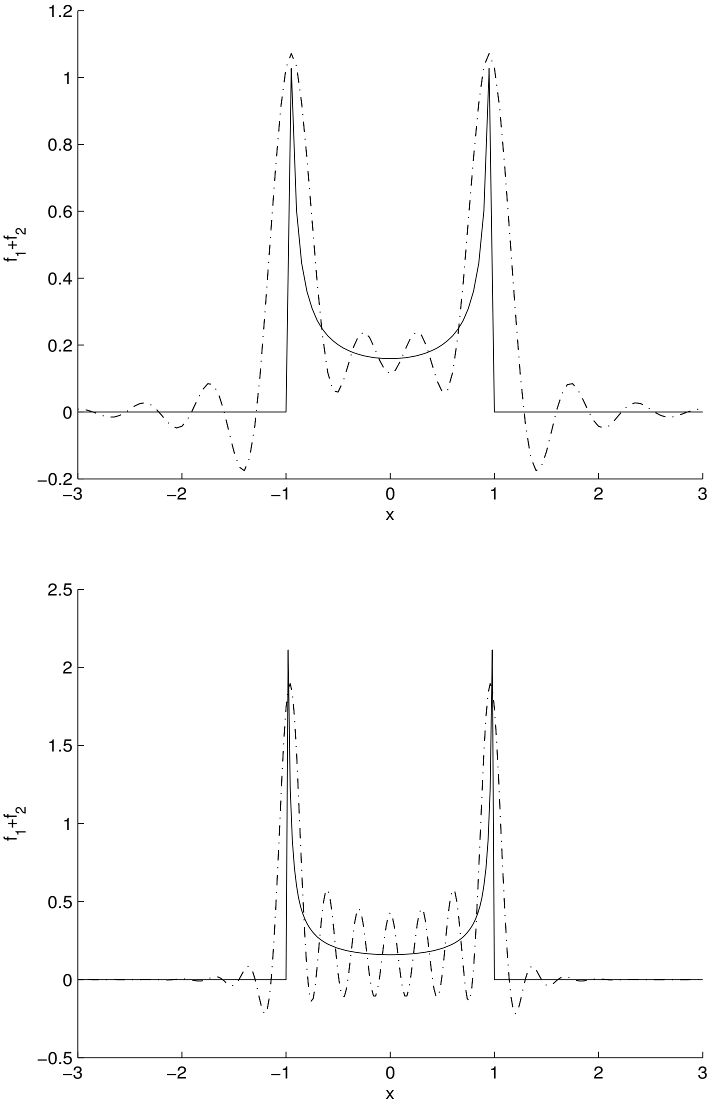


Figure 5.10: Case 2, $\delta < 1$; Numerical (dash-dot line) and equilibrium solution (solid line) to the summation of states $f_1 + f_2$; top: $\alpha = 0.7$, bottom: $\alpha = 1.4$; $N = 100$.

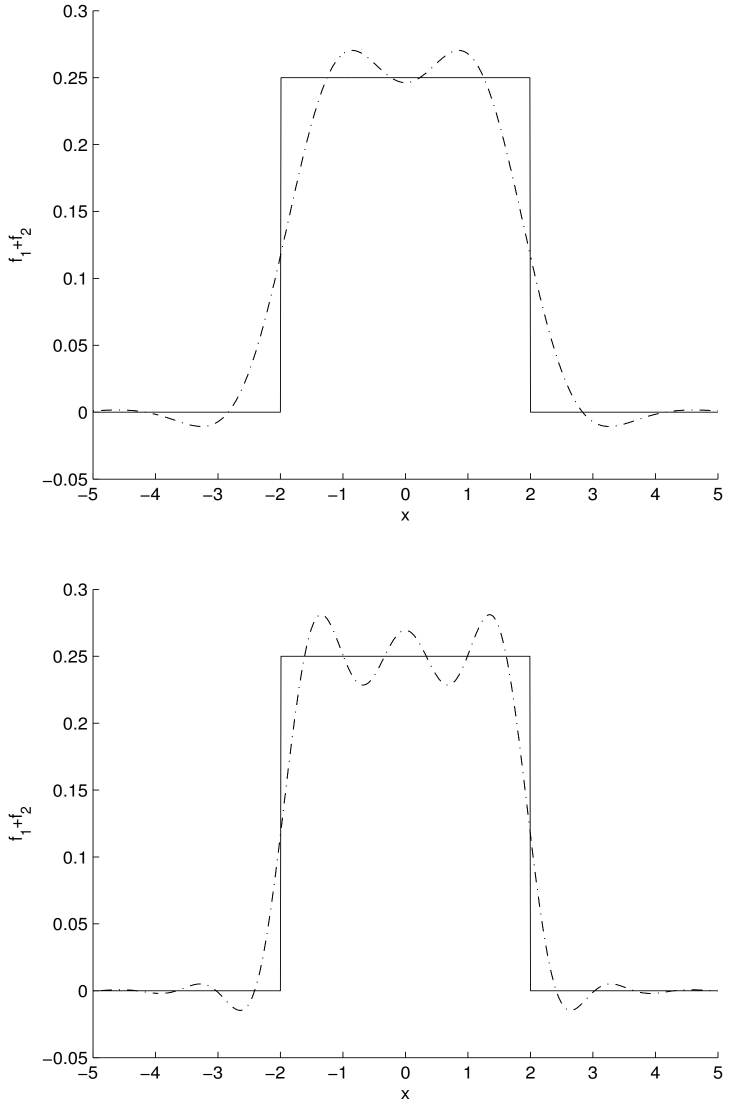


Figure 5.11: Case 3, $\delta = 1$; Numerical (dash-dot line) and equilibrium solution (solid line) to the summation of states $f_1 + f_2$; top: $N = 10$, bottom: $N = 50$; $\alpha = 0.5$.

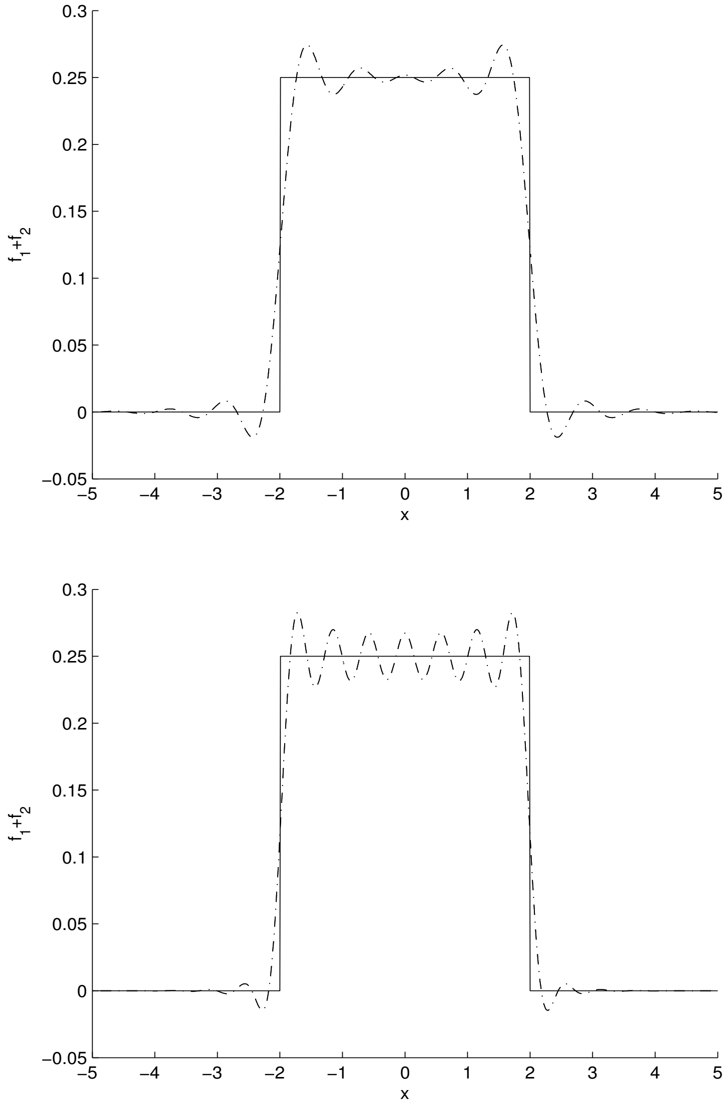


Figure 5.12: Case 3, $\delta = 1$; Numerical (dash-dot line) and equilibrium solution (solid line) to the summation of states $f_1 + f_2$; top: $\alpha = 0.5$, bottom: $\alpha = 0.8$; $N = 100$.

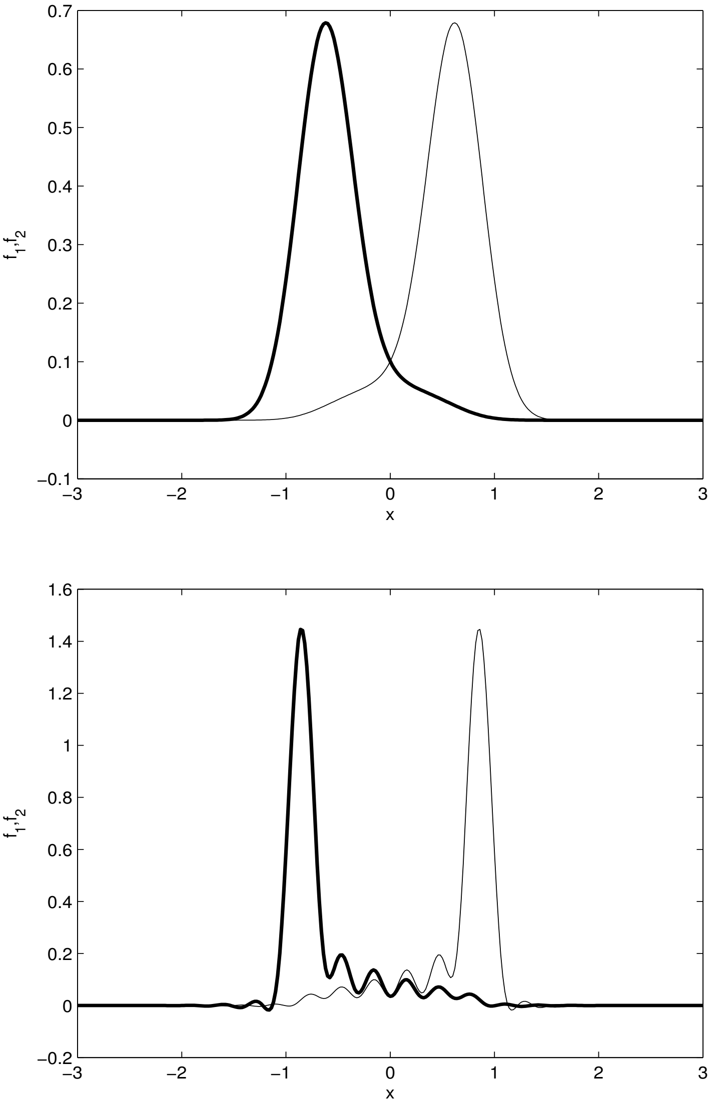


Figure 5.13: Case 2, $\delta < 1$; Numerical solutions at $t = 1$ (top) and $t = 2$ (bottom); $\alpha = 1.4$ and $N = 100$. Thin lines correspond to f_1 and thick lines correspond to f_2.

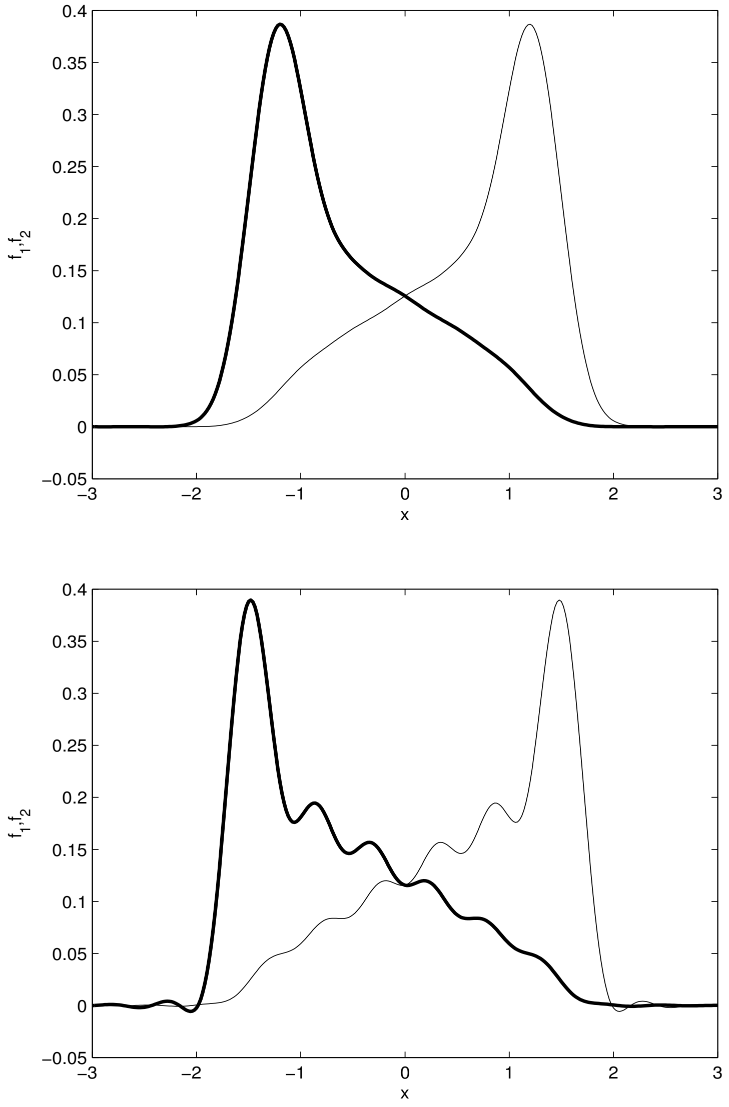


Figure 5.14: Case 3, $\delta = 1$; Numerical solutions at $t = 2$ (top) and $t = 3$ (bottom); $\alpha = 0.8$ and $N = 100$. Thin lines correspond to f_1 and thick lines correspond to f_2.

Chapter 6

Fokker-Planck optimality systems

6.1 Definition of optimal control problems

Optimal control problems consist of the mathematical models in which the underlying task is to transfer the state of a dynamical system from a given initial position into a desired terminal condition. Naturally, there are always practical constraints that are imposed by a particular situation. Nevertheless, there exists generally freedom in the choice of the controls over time to achieve a desired objective. This leads to optimization problems. In some cases, problems are naturally associated with an objective function to be minimized or maximized. However, we also encounter problems in which there is no such choice, and imposing a criterion may simply be a means to generate procedures that allow one to come up with a reasonable solution to the underlying problem. Therefore, the problem of transferring the state of a dynamical system from a given initial condition into a set of desired terminal conditions, while at the same time minimizing some objective associated with the motion, and possibly a penalty on the terminal state, is a most natural one. These belong to the general type of problems that are analyzed with the tools and techniques of optimal control theory. In this chapter, we study two infinite-dimensional optimal control problem.

6.2 Derivation of FP optimality systems

6.2.1 Optimal control of the FP equation for Itō processes

We formulate the problem to determine a control $u \in \mathbb{R}^\ell$ such that starting with an initial distribution ρ the probability density f satisfying the FP equation

$$\partial_t f(x,t) - \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}\left(a_{ij}(x,t)\, f(x,t)\right) + \sum_{i=1}^{d} \partial_{x_i}\left(b_i(x,t;u)\, f(x,t)\right) = 0$$

evolves towards a desired target probability density $f_d(x,t)$ at time $t = T$. This objective can be formulated by the following tracking functional

$$J(f,u) := \frac{1}{2}\|f(\cdot,T) - f_d(\cdot,T)\|^2_{w_\alpha} + \frac{\nu}{2}|u|^2, \tag{6.1}$$

where $|u|^2 = u_1^2 + \ldots + u_\ell^2$, and $\nu > 0$ is a constant. With $\| \cdot \|^2_{w_\alpha}$, we denote the following

$$\|v\|^2_{w_\alpha} = \int_\Omega v(x)^2 w_\alpha(x) dx,$$

where $w_\alpha(x) = \exp(\alpha^2 x^2)$ is a weight function, and α must be appropriately chosen.

The optimal control problem to find u that minimizes the objective J subject to the constraint given by the FP equation is formulated by the following

$$\min \frac{1}{2}\|f(\cdot,T) - f_d(\cdot,T)\|^2_{w_\alpha} + \frac{\nu}{2}|u|^2 \tag{6.2}$$

$$\partial_t f(x,t) - \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}\left(a_{ij}(x,t)\, f(x,t)\right) + \sum_{i=1}^{d} \partial_{x_i}\left(b_i(x,t;u)\, f(x,t)\right) = 0 \tag{6.3}$$

$$f(x,0) = \rho(x). \tag{6.4}$$

Notice that for a given control function u, Theorem 2 states that the solution of the FP model (6.3)-(6.4) is uniquely determined. We denote this dependence by $f = f(u)$ and one can prove that the mapping $u \to f(u)$ is twice differentiable [72]. Therefore, we can introduce the so-called reduced cost functional $\hat{J}$ given by

$$\hat{J}(u) = J(f(u),u). \tag{6.5}$$

Correspondingly, a local minimum u^* of $\hat{J}$ is characterized by $\hat{J}'(u^*;\delta u) = 0$ for all $\delta u \in \mathbb{R}^\ell$. h

To characterize the solution to our optimization problem, we consider the Lagrange formalism and formulate the first-order optimality conditions. Consider the Lagrange functional

$$\begin{aligned} L(f,u,p) &= J(f,u) \\ &+ \int_\Omega \int_0^T \left(\partial_t f - \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}\left(a_{ij}\, f\right) + \sum_{i=1}^{d} \partial_{x_i}\left(b_i(u)\, f\right)\right) p\, w_\alpha\, dx\, dt, \end{aligned}$$

where $p = p(x,t)$ represents the Lagrange multiplier. The first-order optimality conditions for our FP optimal control problem are formally derived by equating to zero the Frechét derivatives of the Lagrange function with respect to the set of variables (f,u,p); see, e.g., [20, 72]. The optimality conditions result in the following optimality system. We have

$$\begin{aligned} \partial_t f - \tfrac{1}{2}\sum_{i,j=1}^d \partial^2_{x_i x_j}(a_{ij}\, f) + \sum_{i=1}^d \partial_{x_i}(b_i(u)\, f) = 0 \quad & \text{in } Q,\\ \text{(state equation)} \quad & \\ f(x,0) = \rho(x) \quad & \text{in } \Omega,\\ \text{(initial condition)} \quad & \\ -\partial_t(p\, w_\alpha) - \tfrac{1}{2}\sum_{i,j=1}^d a_{ij}\, \partial^2_{x_i x_j}(p\, w_\alpha) - \sum_{i=1}^d b_i(u)\, \partial_{x_i}(p\, w_\alpha) = 0 \quad & \text{in } Q,\\ \text{(adjoint equation)} \quad & \\ -p(x,T) = f(x,T) - f_d(x,T) \quad & \text{in } \Omega,\\ \text{(terminal condition)} \quad & \\ \nu\, u_l + \textstyle\int_0^T\!\int_\Omega \left(\sum_{i=1}^d \partial_{x_i}(\tfrac{\partial b_i}{\partial u_l}\, f)\right) p\, w_\alpha\, dx\, dt = 0 \quad \text{in } Q, \quad l = 1,\dots,\ell \quad & \\ \text{(optimality equations)} \quad & \end{aligned} \tag{6.6}$$

Notice that the state variable evolves forward in time and the adjoint variable evolves backwards in time. We remark that the FP equation is a particular instance of the forward Kolmogorov equation and the adjoint equation resembles the backward Kolmogorov equation.

It should appear [6, 7] clearly that the lth component of the reduced gradient $\nabla \hat{J}$ is given by

$$(\nabla \hat{J})_l = \nu\, u_l + \int_0^T\!\!\int_\Omega \left(\sum_{i=1}^d \partial_{x_i}\left(\frac{\partial b_i}{\partial u_l}\, f\right)\right) p\, w_\alpha\, dx\, dt, \qquad l = 1,\dots,\ell, \tag{6.7}$$

where $p = p(u)$ is the solution of the adjoint equation for the given $f(u)$.

Notice that the optimization problem given by (6.2)-(6.4) represents a bilinear control problem where the dependence of the state f on the control u is nonlinear and the corresponding optimization problem is nonconvex. However, standard arguments [6, 7, 20, 72, 104] allow to prove existence of optimal solutions of the open-loop control in $(0,T)$.

6.2.2 Optimal control of the FP equation for PDP processes

We introduce in

$$\partial_t f_s(x,t) + \partial_x(A_s(x) f_s(x,t)) = \sum_{j=1}^S \mathcal{Q}_{sj} f_j(x,t), \qquad s \in \mathbb{S},$$

a control mechanism in the deterministic dynamics of a PDP process, and consider $A_s(x,u_s)$ in (Ω, U_s), where $U_s \subset \mathbb{R}^l$, $s \in \mathbb{S}$, are closed compact sets. We assume that for a given state (x,s) of the system, admissible open-loop control functions $u_s(t) : [0,T) \to U_s$, $s \in \mathbb{S}$, exist and are continuous in the interval $[0,T)$. Further, we assume that $A_s(x,u_s)$, $s \in \mathbb{S}$, are Lipschitz continuous and differentiable in the set (Ω, U_s) so that the differential system

$$\frac{d}{dt} X(t) = A_{\mathscr{S}(t)}(X, u_{\mathscr{S}(t)}(t)), \quad t \in [0,T),$$

has a unique solution. In the following, we denote with $f = (f_s)_{s=1}^S$ and $u = (u_s)_{s=1}^S$. Notice that the control strategy has a bilinear structure.

We consider the problem to find optimal controls u_s, $s \in \mathbb{S}$, such that the solution to the FP model

$$\partial_t f_s(x,t) + \partial_x (A_s(x,u_s) f_s(x,t)) = \sum_{j=1}^S \mathcal{Q}_{sj} f_j(x,t), \qquad s \in \mathbb{S}, \tag{6.8}$$

subject to

$$f_s(x,0) = f_s^0(x), \qquad s \in \mathbb{S}. \tag{6.9}$$

minimizes the following cost functional

$$J(f,u) := \frac{1}{2} \sum_{s=1}^S \|f_s(\cdot,T) - f_s^T(\cdot)\|_{w_\alpha}^2 + \frac{\nu}{2} \sum_{s=1}^S |u_s|_U^2, \tag{6.10}$$

where $f^T = (f_1^T, \cdots, f_S^T) \in C_0^\infty(\mathbb{R}, \mathbb{R}^S)$ is a vector of given functions with trace zero that represents a desired target PDF at time T, $\nu > 0$ is the weight of the cost of the control, and $|.|_U$ denotes a norm in the space of the controls. It is assumed that $J(f,u)$ is twice Frechét-differentiable and that the second Frechét derivative J'' is locally Lipschitz-continuous. This objective models the requirement that the PDF of the PDP at final time, $f_s(\cdot,T)$, approaches as close as possible the desired target f_s^T. We choose the following weighted L^2-norm

$$\|v\|_{w_\alpha}^2 = \int_{\mathbb{R}} v(x)^2 w_\alpha(x) dx,$$

where $w_\alpha(x) = \exp(\alpha^2 x^2)$ is a weight function and α is the scaling factor. In compact form, we have the following optimal control problem

$$\min_{u \in U} J(f,u), \ (f,u) \ \text{subject to (6.8)} - \text{(6.9)}, \tag{6.11}$$

where $U = U_1 \times \cdots \times U_S$. This is an infinite-dimensional constrained minimization problem [20], whose solution can be characterized by the corresponding first-order optimality system. In order to derive this system, we introduce the following Lagrange function

$$\begin{aligned} L(f,u,p) &= J(f,u) \\ &+ \sum_{s=1}^S \int_0^T \int_{\mathbb{R}} \left(\partial_t f_s + \partial_x (A_s(u_s) f_s) - \sum_{j=1}^S \mathcal{Q}_{sj} f_j \right) p_s \, w_\alpha \, dx \, dt, \end{aligned}$$

where $p = (p_s)_{s=1}^S$ is the vector of adjoint PDF variables. The first-order optimality conditions for the optimal control problem (6.11) are formally derived by equating to zero the Frechét derivatives of the Lagrange function with respect to the set of variables (f_s, u_s, p_s); see, e.g., [20]. We obtain the following optimality system. For $s \in \mathbb{S}$, we have

$$\begin{aligned}
\partial_t f_s + \partial_x (A_s f_s) = \textstyle\sum_{j=1}^{S} \mathcal{Q}_{sj} f_j \quad & \text{in } \mathbb{R} \times (0,T), \quad \text{(state equation)}\\
f_s(x,0) = f_s^0(x) \quad & \text{in } \mathbb{R}, \quad \text{(initial condition)}\\
-\partial_t (p_s w_\alpha) - A_s \partial_x (p_s w_\alpha) = \textstyle\sum_{j=1}^{S} \mathcal{Q}_{js} p_j w_\alpha \quad & \text{in } \mathbb{R} \times (0,T), \quad \text{(adjoint equation)}\\
-p_s(x,T) = f_s(x,T) - f_s^T(x) \quad & \text{in } \mathbb{R}, \quad \text{(terminal condition)}\\
\nu\, u_s + \textstyle\int_{\mathbb{R}} \partial_x (\partial_{u_s} A_s f_s)\, p_s w_\alpha \, dx = 0 \quad & \text{in } (0,T). \quad \text{(optimality equation)}
\end{aligned} \tag{6.12}$$

Note that the adjoint system describes the backward evolution of the adjoint variables, starting with the terminal condition at $t = T$, and it is strictly hyperbolic as the forward equation. Hence assuming $f \in C^\infty(\mathbb{R} \times (0,T))$ and a desired smooth target $f^T \in C^\infty(\mathbb{R}, \mathbb{R}^S)$, we have $p \in C^\infty(\mathbb{R} \times (0,T), \mathbb{R}^S)$. Based on these regularity results and on the convexity and differentiability of the cost functional and on the form of the optimality condition, standard arguments apply to prove existence of optimal solutions; see, e.g., [20, 28, 59, 72, 78].

In the case that constant controls in $[0,T]$ are required, we have the following optimality condition

$$\nu\, u_s + \int_0^T \int_{\mathbb{R}} \partial_x (\partial_{u_s} A_s f_s)\, p_s w_\alpha \, dx\, dt = 0.$$

We mention this case because in [8] a model predictive control strategy is investigated where on each time window constant controls are considered. For this reason, the error estimate focuses on constant controls.

Chapter 7

Hermite spectral discretization of FP optimality systems

This chapter aims to introduce and analyse Hermite spectral discretization as an appropriate discretization scheme in the field of optimal control problems involving parabolic and hyperbolic FP equations in unbounded domains with bilinear control structure.

7.1 A FP optimality system for an Itō stochastic process

We consider a FP control problem corresponding to a representative stochastic process given by the Ornstein-Uhlenbeck process, and for simplicity we focus on an one-dimensional setting, $d = 1$, in which the drift function b is linear and the diffusion coefficient a is constant. We have $b(x, t; u) = \gamma x + u$ and $a(x, t) = 2c$, where $\gamma < 0$, u and $c > 0$ are constants. In this case, the optimality system is given by

$$\begin{aligned} \partial_t f(x,t) - c\partial_{xx} f(x,t) + \partial_x \left((\gamma x + u) f(x,t) \right) = 0, &\quad \text{in}\, Q, \\ f(x,0) = \rho(x), &\quad \text{in}\, \Omega, \\ -\partial_t (p(x,t) w_\alpha(x)) - c\partial_{xx}(p(x,t)w_\alpha(x)) - (\gamma x + u)\partial_x (p(x,t) w_\alpha(x)) = 0, &\quad \text{in}\, Q, \\ -p(x,T) = f(x,T) - f_d(x,T), &\quad \text{in}\, \Omega, \\ \nu u + \int_0^T \int_{\mathbb{R}} \partial_x f(x,t) p(x,t) w_\alpha \, dx\, dt = 0, &\quad \text{in}\, Q, \end{aligned}$$

where $w_\alpha(x) = \exp(\alpha^2 x^2)$ is a weight function and α is the scaling factor.

The state and adjoint variables are approximated in the space of Hermite functions as follows

$$f(x,t) = \sum_{n=0}^{\infty} \hat{f}_n(t) \tilde{\mathrm{H}}_n(x), \qquad p(x,t) = \sum_{n=0}^{\infty} \hat{p}_n(t) \tilde{\mathrm{H}}_n(x).$$

For the adjoint equation, we note that the following approximation

$$p(x,t) = \sum_{n=0}^{\infty} \hat{p}_n(t)\tilde{\mathrm{H}}_n(x),$$

is equivalent to

$$p(x,t)w_\alpha(x) = \sum_{n=0}^{\infty} \frac{1}{\sqrt{2^n n!}}\hat{p}_n(t)H_n(\alpha x).$$

The initial data $f(x,0) = \rho$ is also represented in the Hermite functions space by $\rho(x) = \sum_{n=0}^{\infty} \hat{f}_n^0 \tilde{\mathrm{H}}_n(x)$, where

$$\hat{f}_n^0 = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} \rho(x)\tilde{\mathrm{H}}_n(x)w_\alpha(x)dx, \quad n \geq 0.$$

After calculating the numerical solution of the forward equation, we have $f(x,T) = \sum_{n=0}^{\infty} \hat{f}_n(T)\tilde{\mathrm{H}}_n(x)$. Since $p(x,T) = f_d(x,T) - f(x,T)$, the terminal condition for the adjoint variable p can be approximated by $p(x,T) = \sum_{n=0}^{\infty} \hat{p}_{T,n}\tilde{\mathrm{H}}_n(x)$, where

$$\hat{p}_{T,n} = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} f_d(x,T)\tilde{\mathrm{H}}_n(x)w_\alpha(x)dx - \hat{f}_n(T), \quad n \geq 0.$$

Introducing the Hermite expansions for f and p into the state and adjoint equations, for $n \geq 0$ we have

$$\frac{d}{dt}\hat{f}_n(t) = n\gamma\hat{f}_n(t) + \alpha u\sqrt{2n}\hat{f}_{n-1}(t) + (\gamma + 2\alpha^2 c)\sqrt{n(n-1)}\hat{f}_{n-2}(t), \tag{7.1}$$

with $\hat{f}_{-1} = 0$, $\hat{f}_{-2} = 0$, and

$$-\frac{d}{dt}\hat{p}_n(t) = n\gamma\hat{p}_n(t) + \alpha u\sqrt{2(n+1)}\hat{p}_{n+1}(t) + (\gamma + 2\alpha^2 c)\sqrt{(n+2)(n+1)}\hat{p}_{n+2}(t) \tag{7.2}$$

The equations (7.1)-(7.2) represent two infinite systems of ODEs. These systems are truncated by considering the approximations

$$[\hat{f}_{\Delta,0}(t), \hat{f}_{\Delta,1}(t), \cdots, \hat{f}_{\Delta,N}(t)] \approx [\hat{f}_0(t), \hat{f}_1(t), \cdots],$$

and

$$[\hat{p}_{\Delta,0}(t), \hat{p}_{\Delta,1}(t), \cdots, \hat{p}_{\Delta,N}(t)] \approx [\hat{p}_0(t), \hat{p}_1(t), \cdots].$$

Therefore, the systems of ODEs which we solve are as follows

$$\begin{aligned} \tfrac{d}{dt}\hat{f}_{\Delta,n}(t) = n\gamma\hat{f}_{\Delta,n}(t) + \alpha u\sqrt{2n}\hat{f}_{\Delta,n-1}(t) + (\gamma + 2\alpha^2 c)\sqrt{n(n-1)}\hat{f}_{\Delta,n-2}(t), \\ \hat{f}_{\Delta,n}(0) = \hat{\rho}_n, \end{aligned} \tag{7.3}$$

and

$$\begin{aligned} -\tfrac{d}{dt}\hat{p}_{\Delta,n}(t) = n\gamma\hat{p}_{\Delta,n}(t) + \alpha u\sqrt{2(n+1)}\hat{p}_{\Delta,n+1}(t) + (\gamma + 2\alpha^2 c)\sqrt{(n+2)(n+1)}\hat{p}_{\Delta,n+2}(t), \\ \hat{p}_{\Delta,n}(T) = \hat{p}_{T,n}, \end{aligned} \tag{7.4}$$

for $0 \leq n \leq N, 0 \leq t \leq T$, with $\hat{f}_{\Delta,i} = 0$ and $\hat{p}_{\Delta,N-i} = 0$, $i = -1, -2$. Defining $\tau = T - t$ and $\hat{q}_n(t) = \hat{p}_n(\tau)$, the last equation is equivalent to

$$\begin{aligned} \tfrac{d}{dt}\hat{q}_{\Delta,n}(t) = n\gamma\hat{q}_{\Delta,n}(t) + \alpha u\sqrt{2(n+1)}\hat{q}_{\Delta,n+1}(t) + (\gamma + 2\alpha^2 c)\sqrt{(n+2)(n+1)}\hat{q}_{\Delta,n+2}(t), \\ \hat{q}_{\Delta,n}(0) = \hat{p}_{T^k,n}. \end{aligned} \tag{7.5}$$

The systems (7.3) and (7.5) can be written in the following matrix form,

$$\frac{d\hat{f}_\Delta}{dt} = M_f \hat{f}_\Delta, \tag{7.6}$$

and

$$\frac{d\hat{q}_\Delta}{dt} = M_q \hat{q}_\Delta, \tag{7.7}$$

where

$$\begin{aligned} \hat{f}_\Delta &= [\hat{f}_{\Delta,0}(t), \hat{f}_{\Delta,1}(t), \cdots, \hat{f}_{\Delta,N}(t)]^T, \\ \hat{q}_\Delta &= [\hat{q}_{\Delta,0}(t), \hat{q}_{\Delta,1}(t), \cdots, \hat{q}_{\Delta,N}(t)]^T, \end{aligned}$$

and M_f and M_q are two $(N+1) \times (N+1)$ three-diagonal matrices with the elements

$$(M_f)_{ij} = \begin{cases} n\gamma, & i = j, \\ \alpha u\sqrt{2n}, & i - j = 1, \\ (\gamma + 2\alpha^2 c)\sqrt{n(n-1)}, & i - j = 2, \\ 0, & \text{otherwise}, \end{cases} \qquad 1 \leq i, j \leq N+1,$$

$$(M_q)_{ij} = \begin{cases} n\gamma, & j = i, \\ \alpha u\sqrt{2(n+1)}, & j - i = 1, \\ (\gamma + 2\alpha^2 c)\sqrt{(n+2)(n+1)}, & j - i = 2, \\ 0, & \text{otherwise}, \end{cases} \qquad 1 \leq i, j \leq N+1,$$

where $n = i - 1$. Notice that, the first row in M_f and also the first column in M_q are zero.

Once we have calculated $\hat{f}_\Delta$ and $\hat{q}_\Delta$ by

$$\hat{f}_\Delta(t) = \exp(M_f t)\, \hat{f}^0_\Delta, \quad \text{and} \quad \hat{q}_\Delta(t) = \exp(M_q t)\, \hat{q}^0_\Delta,$$

the optimal control variable u can be computed. Representing the approximated solutions by

$$f_\Delta(x,t) = \sum_{n=0}^{N} \hat{f}_{\Delta,n}(t)\tilde{\mathrm{H}}_n(x), \quad \text{and} \quad p_\Delta(x,t) = \sum_{n=0}^{N} \hat{p}_{\Delta,n}(t)\tilde{\mathrm{H}}_n(x),$$

we have

$$\begin{aligned}
\int_{\mathbb{R}}(\partial_x f_\Delta)p_\Delta w_\alpha dx &= \int_{\mathbb{R}}\left(\sum_{n=0}^{N}\hat{f}_{\Delta,n}\frac{d}{dx}\tilde{\mathrm{H}}_n(x)\right)\left(\sum_{n=0}^{N}\hat{p}_{\Delta,n}\tilde{\mathrm{H}}_n(x)\right)w_\alpha dx\\
&= -\alpha\int_{\mathbb{R}}\left(\sum_{n=0}^{N}\sqrt{2(n+1)}\hat{f}_{\Delta,n}\tilde{\mathrm{H}}_{n+1}(x)\right)\left(\sum_{n=0}^{N}\hat{p}_{\Delta,n}\tilde{\mathrm{H}}_n(x)\right)w_\alpha dx\\
&= -\alpha\sum_{n=0}^{N}\sum_{k=0}^{N}\sqrt{2(n+1)}\hat{f}_{\Delta,n}\hat{p}_{\Delta,k}\int_{\mathbb{R}}\tilde{\mathrm{H}}_{n+1}(x)\tilde{\mathrm{H}}_k(x)w_\alpha dx\\
&= -\alpha\sum_{n=0}^{N-1}\sqrt{2(n+1)}\hat{f}_{\Delta,n}\hat{p}_{\Delta,n+1}\frac{\sqrt{\pi}}{\alpha}\\
&= -\sum_{n=0}^{N-1}\sqrt{2\pi(n+1)}\hat{f}_{\Delta,n}\hat{p}_{\Delta,n+1}.
\end{aligned}$$

Then $\nu u + \int_0^T \int_{\mathbb{R}} \partial_x f(x,t)\, p(x,t)\, w_\alpha \, dx\, dt = 0$ gives the following

$$\begin{aligned}
u_\Delta &= -\frac{1}{\nu}\left\langle \partial_x f_\Delta, p_\Delta\right\rangle_{w_\alpha} = -\frac{1}{\nu}\int_0^T\int_{\mathbb{R}}(\partial_x f_\Delta)p_\Delta w_\alpha dx\, dt\\
&= \frac{1}{\nu}\sum_{n=0}^{N-1}\sqrt{2\pi(n+1)}\int_0^T \hat{f}_{\Delta,n}\hat{p}_{\Delta,n+1}\, dt.
\end{aligned}$$

7.1.1 Convergence analysis

We recall that substituting the Hermite expansion into the FP control system results in two infinite systems of linear ODEs. Corresponding to each system, there is a matrix M_∞, which is lower triangular for the state equation and upper triangular for the adjoint equation. To have a practical scheme, we have to truncate these matrices, or equivalently, consider some truncated systems of ODEs. However, this truncation is a source of error in our discretization scheme. In the following, we investigate the influence of this error on the accuracy of our approximation method. Let $\|.\|_2$ be the Euclidean norm in $\mathbb{R}^{N+1}$.

Lemma 11. *Assuming $f(\cdot,t), f_d(\cdot,T) \in V_N$ for any $t \in [0,T]$, and N is sufficiently large so that there is no error in the spectral representation of the initial data, then*

$$\|\hat{f}_N - \hat{f}_\Delta\|_2 = 0, \quad \textit{and} \quad \|\hat{p}_N - \hat{p}_\Delta\|_2 = 0.$$

That is, there will be no error for the truncation of the infinite ODE systems.

Proof. For the forward case, no truncation error appears in calculating the Hermite coefficients $\hat{f}_n$ by solving the finite ODE system (7.6). This is because of the fact that the system (7.1) is uncoupled in the sense that for $m > n$ the value of $\hat{f}_n$ is independent of the value of $\hat{f}_m$. That is, $P_N f(\cdot,t) = f_\Delta(\cdot,t)$ for every $t \in [0,T]$.

The backward case can be analysed following the procedure proposed in [25]. Consider M_∞ as the representing matrix for the ODE system transformed of the adjoint equation, and M the corresponding truncated matrix. The matrix M is obtained from M_∞ by removing all rows and columns with index larger than $N+1$. We can write

$$\hat{q} = e^{M_\infty t}\hat{q^0} = \left(\sum_{j=0}^{\infty} \frac{t^j M_\infty^j}{j!}\right)\hat{q^0} = \sum_{j=0}^{\infty} \frac{t^j}{j!}(M_\infty^j \hat{q^0}).$$

That is, for $n \geq 1$ we have

$$\hat{q}_{n-1} = \sum_{j=0}^{\infty} \frac{t^j}{j!}(M_\infty^j \hat{q^0})_n = \sum_{j=0}^{\infty} \frac{t^j}{j!} b_n^j,$$

in which $b_n^j = (M_\infty^j \hat{q^0})_n$, and the notation $(v)_n$ refers to the n-th component of the vector v. Since we have assumed that $f(\cdot,t), f_d(\cdot,T) \in V_N$, we have $(\hat{q^0})_n = 0$ for $n > N+1$. Noting that M is an upper triangular matrix, it follows that $(M_\infty \hat{q^0})_n = 0$ for $n > N+1$. Therefore, for $j \geq 1$,

$$b_n^j = \sum_{k=1}^{N+1} (M_\infty^j)_{nk}(\hat{q^0})_k = \sum_{k=1}^{N+1} (M_\infty)_{nk}(M_\infty^{j-1}\hat{q^0})_k = \sum_{k=1}^{N+1} (M_\infty)_{nk} b_k^{j-1}.$$

Similarly $\hat{f}_{\Delta,n-1} = \sum_{j=0}^{\infty} \frac{t^j}{j!} b_{\Delta,n}^j$, with

$$b_{\Delta,n}^j = (M^j \hat{f_\Delta^0})_n = \begin{cases} \sum_{k=1}^{N+1} (M)_{nk} b_{\Delta,k}^{j-1}, & j \geq 1, n \leq N+1, \\ (\hat{f_\Delta^0})_n, & j = 0, n \leq N+1, \\ 0, & n > N+1. \end{cases}$$

Therefore we have

$$\hat{f}_{n-1} - \hat{f}_{\Delta,n-1} = \sum_{j=0}^{\infty} \frac{t^j}{j!}(b_n^j - b_{\Delta,n}^j), \qquad n = 1,2,\cdots,N+1,$$

and consequently by introducing $\theta_j = \sum_{n=1}^{N+1} |b_n^j - b_{\Delta,n}^j|$,

$$\sum_{n=1}^{N+1} |\hat{f}_{n-1} - \hat{f}_{\Delta,n-1}| \leq \sum_{j=0}^{\infty} \frac{t^j}{j!} \sum_{n=1}^{N+1} |b_n^j - b_{\Delta,n}^j| = \sum_{j=0}^{\infty} \frac{t^j}{j!}\theta_j.$$

Through the following argument, we find an upper bound for θ_j.

$$\begin{aligned} \theta_j &= \sum_{n=1}^{N+1} |b_n^j - b_{\Delta,n}^j| \\ &= \sum_{n=1}^{N+1} |\sum_{k=1}^{N+1} (M_\infty)_{nk} b_k^{j-1} - \sum_{k=1}^{N+1} (M)_{nk} b_{\Delta,k}^{j-1}| \qquad (7.8) \\ &\leq \sum_{n=1}^{N+1}\sum_{k=1}^{N+1} |(M)_{nk}||b_k^{j-1} - b_{\Delta,k}^{j-1}|. \end{aligned}$$

Therefore, we have

$$\theta_j \leq \sum_{k=1}^{N+1} |b_k^{j-1} - b_{\Delta,k}^{j-1}| \sum_{n=1}^{N+1} |(M)_{nk}| \leq cN^2 \sum_{k=1}^{N+1} |b_k^{j-1} - b_{\Delta,k}^{j-1}| = cN^2 \theta_{j-1},$$

for some positive constant c. With some calculations, we see that $\theta_j \leq (cN^2)^j \theta_0$. Since $b_n^0 = \hat{f_n^0}$ and $b_{\Delta,n}^0 = \hat{f_{\Delta,n}^0}$, $\theta_0 = \sum_{n=1}^{N+1} |b_n^0 - b_{\Delta,n}^0| = \sum_{n=0}^{N} |\hat{f_n^0} - \hat{f_{\Delta,n}^0}|$ is the error in the representation of the initial data which is assumed to be zero. Therefore, we have

$$\begin{aligned} \|\hat{q}_N - \hat{q}_\Delta\|_2^2 &= \sum_{n=0}^{N} |\hat{q}_n - \hat{q}_{\Delta,n}|^2 \leq \left(\sum_{n=0}^{N} |\hat{q}_n - \hat{q}_{\Delta,n}| \right)^2 \\ &\leq \left(\sum_{j=0}^{\infty} \frac{t^j}{j!} \theta_j \right)^2 \leq \left(\sum_{j=0}^{\infty} \frac{t^j}{j!} (cN^2)^j \theta_0 \right)^2 = 0, \end{aligned}$$

and thence the desired result. □

To analyze the accuracy of the scheme, we first show that the approximations for the state and adjoint variables are spectrally convergent.

Theorem 15. *If $f, p \in L^\infty(0,T; H^r_{w_\alpha}(\mathbb{R}))$, $r > 1$, then for all $t \in [0,T]$ the following holds*

$$\|f(\cdot,t) - f_\Delta(\cdot,t)\|_{w_\alpha}^2 = O(N^{-r}) \quad \textit{and} \quad \|p(\cdot,t) - p_\Delta(\cdot,t)\|_{w_\alpha}^2 = O(N^{-r})$$

for $f_\Delta(x,t) = \sum_{n=0}^{N} \hat{f}_{\Delta,n}(t) \tilde{H}_n(x)$ and $p_\Delta(x,t) = \sum_{n=0}^{N} \hat{p}_{\Delta,n}(t) \tilde{H}_n(x)$.

Proof. The argument is the same for f and p, so we only discuss the statement for p. We have

$$\begin{aligned} p(x,t) - p_\Delta(x,t) &= \sum_{n=0}^{\infty} \hat{p}_n(t) \tilde{\mathrm{H}}_n(x) - \sum_{n=0}^{N} \hat{p}_{\Delta,n}(t) \tilde{\mathrm{H}}_n(x) \\ &= \sum_{n=0}^{N} \left(\hat{p}_n(t) - \hat{p}_{\Delta,n}(t) \right) \tilde{\mathrm{H}}_n(x) + \sum_{n=N+1}^{\infty} \hat{p}_n(t) \tilde{\mathrm{H}}_n(x). \end{aligned}$$

From Lemma 11 we know that the first term in the last line of the equation above is zero, hence Lemma 7 in Chapter 5 gives us the following bound for the error

$$\begin{aligned} \|p(\cdot,t) - p_\Delta(\cdot,t)\|_{w_\alpha}^2 &= \int_{\mathbb{R}} \left[\sum_{n=N+1}^{\infty} \hat{p}_n(t) \tilde{\mathrm{H}}_n(x) \right]^2 w(x) dx \\ &= \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} |\hat{p}_n(t)|^2 \\ &\leq \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} \frac{\alpha^{2-2r}}{\pi} n^{-r} \|p(\cdot,t)\|_{r,w_\alpha}^2. \end{aligned}$$

Therefore, we have $\|p(\cdot,t) - p_\Delta(\cdot,t)\|_{w_\alpha}^2 = O(N^{-r})$. □

The following lemma provides an appropriate means to show that the Hermite discretization method is stable.

Lemma 12. *Let M be the $(N+1)\times(N+1)$ matrix M_f or M_q, and let $\hat{y}(t)$ be the solution to*

$$\frac{d}{dt}\hat{y}(t) = M\hat{y}, \qquad \hat{y}(0) = \hat{y}_0.$$

Then there exists a constant C_N such that for all $t > 0$

$$\|\hat{y}(t)\|_2 \leq C_N \|\hat{y}_0\|_2.$$

Proof. Since the matrix M is triangular, it has $N+1$ distinct eigenvalues $\lambda_n = n\gamma$, $n = 0, 1, \cdots, N$, which are the diagonal elements of M. Therefore, M is diagonalizable and can be decomposed as $M = S^{-1}DS$, where $D = diag(\lambda_n)_{n=0}^{N}$. Hence, the system of ODEs has the solution

$$\hat{y}(t) = e^{Mt}\hat{y}_0,$$

which implies the following

$$\|\hat{y}(t)\|_2 \leq \|S^{-1}\|_2 \|e^{Dt}\|_2 \|S\|_2 \|\hat{y}_0\|_2.$$

Since $\gamma < 0$, we have $e^{2\lambda_n t} \leq 1$, $n = 0, 1, \cdots, N$, and consequently

$$\|e^{Dt}\|_2 = \sigma_{\max}(e^{Dt}) = \sqrt{\lambda_{\max}(e^{2Dt})} = 1.$$

It is easy to show that the matrices S and S^{-1} have the same structure as the matrix M. That is, they are lower triangular when M is lower triangular, and upper triangular when M is upper triangular. Since S is consist of the eigenvectors of M, it can be constructed in such a way that all diagonal elements are 1. Defining $\tilde{s} := \|S\|_{\max}$ we have

$$\|S\|_2 \leq (N+1)\|S\|_{\max} = (N+1)\tilde{s}.$$

Furthermore, in [69] it is proved that

$$\|S^{-1}\|_\infty \leq (\tilde{s}+1)^N,$$

which results in

$$\|S^{-1}\|_2 \leq \sqrt{N+1}\|S^{-1}\|_\infty = \sqrt{N+1}(\tilde{s}+1)^N.$$

Therefore, we have

$$\|\hat{y}(t)\|_2 \leq C_N \|\hat{y}_0\|_2.$$

where $C_N = (N+1)^{3/2}(\tilde{s}+1)^{N+1}$.

□

Based on Lemma 12, we have the following stability result.

Theorem 16. *There exists a constant C_N such that for all $t > 0$*

$$\|f_\Delta(.,t)\|_{w_\alpha} \leq C_N \|\hat{f}_0\|_2, \quad \text{and} \quad \|p_\Delta(.,t)\|_{w_\alpha} \leq C_N \|\hat{p}_0\|_2.$$

Proof. We only prove the inequality for f_Δ, since the argument is the same for p_Δ. We have

$$\begin{aligned}\|f_\Delta(.,t)\|^2_{w_\alpha} &= \int_{\mathbb{R}} (f_\Delta)^2 w_\alpha(x)dx = \int_{\mathbb{R}} \left(\sum_{n=0}^{N} \hat{f}_{\Delta,n}(t) \tilde{\mathrm{H}}_n(x) \right)^2 w_\alpha(x)dx \\ &= \frac{\sqrt{\pi}}{\alpha} \sum_{n=0}^{N} (\hat{f}_{\Delta,n}(t))^2 = \frac{\sqrt{\pi}}{\alpha} \|\hat{f}_\Delta(t)\|_2^2 \leq C_N \|\hat{f}_0\|_2^2.\end{aligned}$$

□

Now, we investigate the spectral convergence in approximating the control variable.

Theorem 17. *Let $f \in L^\infty(0,T; H^r_{w_\alpha}(\mathbb{R}))$, $r > 2$ and $N \geq 2$. Then for a positive constant c, we have*

$$|u - u_\Delta| \leq c\,T \sum_{n=N}^{\infty} n^{1-r}.$$

Proof. We start from optimality equations in the continuous and discrete form:

$$\begin{aligned}\nu u + \int_0^T \int_{\mathbb{R}} (\partial_x f)\, p\, w_\alpha \, dx\, dt = 0, \\ \nu u_\Delta + \int_0^T \int_{\mathbb{R}} (\partial_x f_\Delta)\, p_\Delta \, w_\alpha \, dx\, dt = 0.\end{aligned}$$

We note that

$$\begin{aligned}\int_{\mathbb{R}} (\partial_x f) p w_\alpha dx &= \int_{\mathbb{R}} \left(\sum_{n=0}^{\infty} \hat{f}_n \partial_x \tilde{\mathrm{H}}_n(x) \right) \left(\sum_{n=0}^{\infty} \hat{p}_n \tilde{\mathrm{H}}_n(x) \right) w_\alpha dx \\ &= -\alpha \int_{\mathbb{R}} \left(\sum_{n=0}^{\infty} \sqrt{2(n+1)} \hat{f}_n \tilde{\mathrm{H}}_{n+1}(x) \right) \left(\sum_{n=0}^{\infty} \hat{p}_n \tilde{\mathrm{H}}_n(x) \right) w_\alpha dx \\ &= -\alpha \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \sqrt{2(n+1)} \hat{f}_n \hat{p}_k \int_{\mathbb{R}} \tilde{\mathrm{H}}_{n+1}(x) \tilde{\mathrm{H}}_k(x) w_\alpha dx \\ &= -\alpha \sum_{n=0}^{\infty} \sqrt{2(n+1)} \hat{f}_n \hat{p}_{n+1} \frac{\sqrt{\pi}}{\alpha} \\ &= -\sum_{n=0}^{\infty} \sqrt{2\pi(n+1)} \hat{f}_n \hat{p}_{n+1}.\end{aligned}$$

Similarly, we have

$$\int_{\mathbb{R}} (\partial_x f_\Delta) p_\Delta w_\alpha dx = -\sum_{n=0}^{N-1} \sqrt{2\pi(n+1)} \hat{f}_{\Delta,n} \hat{p}_{\Delta,n+1}.$$

Therefore,

$$\begin{aligned} -\left(\int_{\mathbb{R}} (\partial_x f) p w_\alpha dx - \int_{\mathbb{R}} (\partial_x f_\Delta) p_\Delta w_\alpha dx \right) &= \sum_{n=0}^{\infty} \sqrt{2\pi(n+1)} \hat{f}_n \hat{p}_{n+1} \\ &- \sum_{n=0}^{N-1} \sqrt{2\pi(n+1)} \hat{f}_{\Delta,n} \hat{p}_{\Delta,n+1} \\ &= \sum_{n=0}^{N-1} \sqrt{2\pi(n+1)} \left(\hat{f}_n \hat{p}_{n+1} - \hat{f}_{\Delta,n} \hat{p}_{\Delta,n+1} \right) \\ &+ \sum_{n=N}^{\infty} \sqrt{2\pi(n+1)} \hat{f}_n \hat{p}_{n+1}. \end{aligned}$$

Noting that

$$\begin{aligned} \hat{f}_n \hat{p}_{n+1} - \hat{f}_{\Delta,n} \hat{p}_{\Delta,n+1} &= \hat{f}_n \hat{p}_{n+1} - \hat{f}_n \hat{p}_{\Delta,n+1} + \hat{f}_n \hat{p}_{\Delta,n+1} - \hat{f}_{\Delta,n} \hat{p}_{\Delta,n+1} \\ &= \hat{f}_n \left(\hat{p}_{n+1} - \hat{p}_{\Delta,n+1} \right) + \hat{p}_{\Delta,n+1} \left(\hat{f}_n - \hat{f}_{\Delta,n} \right), \end{aligned}$$

we can write

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \left| \int_{\mathbb{R}} (\partial_x f) p w_\alpha dx - \int_{\mathbb{R}} (\partial_x f_\Delta) p_\Delta w_\alpha dx \right| &\leq \sum_{n=0}^{N-1} \sqrt{n+1} |\hat{f}_n| . |\hat{p}_{n+1} - \hat{p}_{\Delta,n+1}| \\ &+ \sum_{n=0}^{N-1} \sqrt{n+1} |\hat{p}_{\Delta,n+1}| . |\hat{f}_n - \hat{f}_{\Delta,n}| \\ &+ \sum_{n=N}^{\infty} \sqrt{n+1} |\hat{f}_n| . |\hat{p}_{n+1}|. \end{aligned}$$

From Lemma 11, we have

$$\sum_{n=0}^{N} |\hat{f}_n - \hat{f}_{\Delta,n}|^2 = 0, \quad \text{and} \quad \sum_{n=0}^{N} |\hat{p}_{\Delta,n} - \hat{p}_n|^2 = 0,$$

which implies that $|\hat{f}_n - \hat{f}_{\Delta,n}| = 0$ and $|\hat{p}_{\Delta,n} - \hat{p}_n| = 0$ for $n = 0, 1, \cdots, N$. Hence

$$\frac{1}{\sqrt{2\pi}} \left| \int_{\mathbb{R}} (\partial_x f) p w_\alpha dx - \int_{\mathbb{R}} (\partial_x f_\Delta) p_\Delta w_\alpha dx \right| \leq \sum_{n=N}^{\infty} \sqrt{n+1} |\hat{f}_n| . |\hat{p}_{n+1}|.$$

Assuming $N \geq 2$, Lemma 7 in Chapter 5 gives us

$$\begin{aligned}\sum_{n=N}^{\infty} \sqrt{n+1}|\hat{f}_n|.|\hat{p}_{n+1}| &\leq \sum_{n=N}^{\infty} \sqrt{n+1}(\frac{\alpha^{1-r}}{\sqrt{\pi}})^2 n^{-r/2}(n+1)^{-r/2}\|f\|_{r,w_\alpha}\|p\|_{r,w_\alpha} \\ &\leq c_{fp}\sum_{n=N}^{\infty} n\, n^{-r/2}n^{-r/2} = c_{fp}\sum_{n=N}^{\infty} n^{1-r},\end{aligned}$$

in which $c_{fp} = \frac{\alpha^{2-2r}}{\pi}\|f\|_{r,w_\alpha}\|p\|_{r,w_\alpha}$. Therefore, the desired accuracy estimate for the control variable u can be written as follows

$$|u - u_\Delta| \leq c\,T\sum_{n=N}^{\infty} n^{1-r},$$

where $c = \frac{\sqrt{2\pi}}{\nu}c_{fp}$. □

7.1.2 Numerical experiments

As the first case, Case 1, we set $c = 1$, $\gamma = -1$, $\nu = 0.1$, $T = 1$, and introduce the following initial condition for the FP equation

$$\rho(x) = e^{(-\frac{x^2}{2})}\left(1 + \cos(\frac{\pi}{2}x)\exp(\frac{\pi^2}{8})\right), \quad x \in \mathbb{R}.$$

We consider this case since the exact solution of the FP equation is known, which makes it possible to evaluate the accuracy of the discretization method. The positivity of the FP equation is not considered in this experiment. In order to illustrate the spectral accuracy of the adjoint and the control variables, we insert the desired function

$$f_d(x,t) = e^{(-\frac{x^2}{2})}\left(1 + \cos(\frac{\pi}{2}xe^{-t})\exp(\frac{\pi^2}{8})e^{-2t}\right)$$

into the optimality system, which is the same as the solution of the forward FP equation. with this setting, the exact solution of the optimality system is given by $f = f_d$, $p = 0$, and $u = 0$. We apply our spectral discretization for this optimality system, and obtain very accurate numerical approximations; see Table 7.1 for the norm of the solution errors. We observe that the Hermite spectral method converges spectrally and is very accurate even for small N.

In Case 2, we impose the PDF to follow a desired function which is a Gaussian with a varying mean value. The control variable, then must vary in order to keep the PDF as close as possible to the desired function. Let $\nu = 0.1$,

$$f_d(x,t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x - 2\sin(\pi t/5))^2}{2\sigma^2}\right)$$

with $\sigma = 0.2$, and consider the following setting for the evolution of PDF. The initial PDF is

$$f_0(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

N	$\|f_\Delta - f_{exact}\|_{L^2}$	$\|p_\Delta - p_{exact}\|_{L^2}$	$\vert u_\Delta - u_{exact}\vert$
5	1.6858e-05	5.3832e-15	2.1653e-16
10	6.8641e-10	5.3836e-15	5.9684e-16
15	4.3062e-13	5.3870e-15	2.1858e-15
20	1.0100e-14	5.9588e-15	1.5715e-16

Table 7.1: Case 1: Accurate approximation results of the optimality system for different N.

with $\sigma = 0.5$, and the parameters in the forward equation are $\gamma = -1$, $c = 0.32$. We consider a model predictive control scheme, which is introduced in [7] and presented at the Appendix of this work, to track f_d by the PDF. In this control scheme, we divide the time interval $[0, T]$ into k subintervals, and solve the optimization problem for any time window of size $\Delta t = T/k$. At any time window $(t_k, t_{k+1}]$ an optimal control u imposes the PDF of that window to evolve towards the desired function $f_d(x, t_{k+1})$. While for a given u, the state and the adjoint variables are approximated directly with the Hermite spectral method, we employ the nonlinear conjugate gradient scheme proposed in [7] to evaluate the optimal control u. The final PDF of a window is considered to be the initial solution of the next window. Figure 7.1 along with Table 7.2 show the outcome of this control strategy. In this experiment, $\Delta t = 0.5$, $\alpha = 0.7$ and $N = 50$.

Time interval	u
(0,0.5]	1.1374
(0.5,1]	1.7166
(1,1.5]	2.0767
(1.5,2]	2.1216
(2,2.5]	1.9601
(2.5,3]	1.5307
(3,3.5]	0.9934
(3.5,4]	0.4265
(4,4.5]	0.0019
(4.5,5]	0.0000

Table 7.2: Case 2: The optimal control variable u at different time windows.

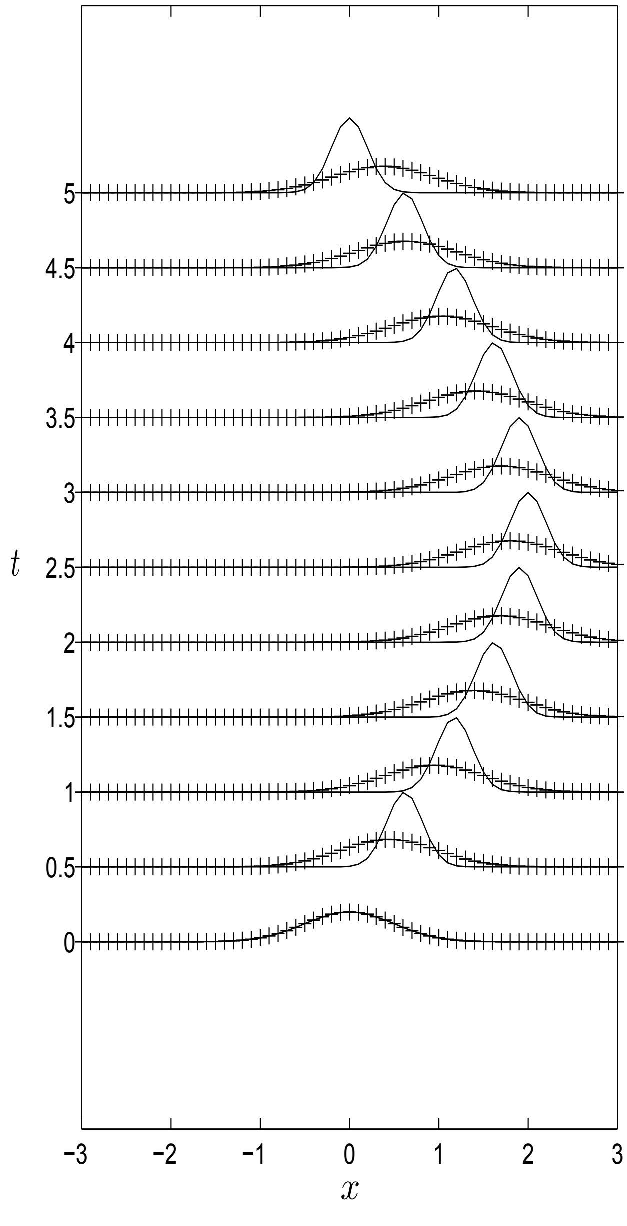


Figure 7.1: Case 2: Approximated solution of FP equation (cross-marks) tracking the desired PDF (solid line) at different time windows.

7.2 A FP optimality system for a PDP Process

We consider the FP optimality system corresponding to a PDP process, which has been derived in Chapter 6, as follows

$$\begin{aligned}
\partial_t f_s + \partial_x(A_s f_s) &= \textstyle\sum_{j=1}^S \mathcal{Q}_{sj} f_j && \text{in } \mathbb{R}\times(0,T), \quad \text{(state equation)}\\
f_s(x,0) &= f_s^0(x) && \text{in } \mathbb{R}, \quad \text{(initial condition)}\\
-\partial_t(p_s w_\alpha) - A_s\partial_x(p_s w_\alpha) &= \textstyle\sum_{j=1}^S \mathcal{Q}_{js} p_j w_\alpha && \text{in } \mathbb{R}\times(0,T), \quad \text{(adjoint equation)}\\
-p_s(x,T) &= f_s(x,T) - f_s^T(x) && \text{in } \mathbb{R}, \quad \text{(terminal condition)}\\
\nu\, u_s + \textstyle\int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_s)\, p_s w_\alpha\, dx &= 0 && \text{in } \mathbb{R}\times(0,T). \quad \text{(optimality equation)}
\end{aligned} \tag{7.9}$$

and focus on the following structure, $A_s(x,u_s) = a_s x + b_s u_s + c_s$, where a_s, b_s and c_s are constants and the u_s are given, $s \in \mathbb{S}$. This choice of the dynamics A_s includes the dissipative process subject to dichotomic noise considered in [8]. In this case, the proposed scheme can be validated by comparison with some exact solutions.

For developing the numerical scheme, we start by discretizing the forward system

$$\partial_t f_s + \partial_x(A_s(x,u_s) f_s) = \sum_{j=1}^S \mathcal{Q}_{sj} f_j, \qquad s \in \mathbb{S}.$$

After replacing A_s, this system becomes

$$\partial_t f_s = -a_s x\, \partial_x f_s - (b_s u_s + c_s)\partial_x f_s + (\mathcal{Q}_{ss} - a_s) f_s + \sum_{\substack{j=1\\ j\neq s}}^S \mathcal{Q}_{sj} f_j, \qquad s \in \mathbb{S}. \tag{7.10}$$

We denote the approximation of $f_s(\cdot,t)$ with $f_\Delta^s(\cdot,t) \in V_N$, where

$$f_\Delta^s(x,t) = \sum_{n=0}^N \hat{f}_n^s(t)\tilde{\mathrm{H}}_n(x). \tag{7.11}$$

We insert (7.11) into (7.10) and obtain

$$\frac{d}{dt}\hat{f}_n^s(t) = d_n^s \hat{f}_n^s(t) + l_n^s \hat{f}_{n-1}^s(t) + h_n^s \hat{f}_{n-2}^s(t) + \sum_{\substack{j=1\\ j\neq s}}^S k_j^s \hat{f}_n^j(t), \quad s \in \mathbb{S}, \tag{7.12}$$

where $d_n^s = \mathcal{Q}_{ss} + a_s n$, $l_n^s = \alpha\sqrt{2n}(b_s u_s + c_s)$, $h_n^s = a_s\sqrt{n(n-1)}$, $k_j^s = \mathcal{Q}_{sj}$ and $\hat{f}_{-1}^s = \hat{f}_{-2}^s = 0$. Equation (7.12) is valid for $0 \leq n \leq N$ and gives the following system of ODEs

$$\frac{d\hat{f}_\Delta(t)}{dt} = M_f(t)\hat{f}_\Delta(t), \tag{7.13}$$

where

$$\hat{f}_\Delta = [\hat{f}_0^1, \cdots, \hat{f}_N^1, \hat{f}_0^2, \cdots, \hat{f}_N^2, \cdots, \hat{f}_0^S, \cdots, \hat{f}_N^S]^T.$$

The matrix M_f is a $S(N+1) \times S(N+1)$ sparse matrix as follows

$$M_f = \left[\begin{array}{cccccccccccccccc}
d_0^1 & & & & & k_2^1 & & & & & & k_S^1 & & & & \\
l_1^1 & d_1^1 & & & & & k_2^1 & & & & & & k_S^1 & & & \\
h_2^1 & l_2^1 & d_2^1 & & & & & k_2^1 & & & \cdots & & & k_S^1 & & \\
 & \ddots & \ddots & \ddots & & & & & \ddots & & & & & & \ddots & \\
 & & h_N^1 & l_N^1 & d_N^1 & & & & & k_2^1 & & & & & & k_S^1 \\
k_1^2 & & & & & d_0^2 & & & & & & k_S^2 & & & & \\
 & k_1^2 & & & & l_1^2 & d_1^2 & & & & & & k_S^2 & & & \\
 & & k_1^2 & & & h_2^2 & l_2^2 & d_2^2 & & & \cdots & & & k_S^2 & & \\
 & & & \ddots & & & \ddots & \ddots & \ddots & & & & & & \ddots & \\
 & & & & k_1^2 & & & h_N^2 & l_N^2 & d_N^2 & & & & & & k_S^2 \\
 & & \vdots & & & & & \vdots & & & \ddots & & & \vdots & & \\
k_1^S & & & & & k_2^S & & & & & & d_0^S & & & & \\
 & k_1^S & & & & & k_2^S & & & & & l_1^S & d_1^S & & & \\
 & & k_1^S & & & & & k_2^S & & & \cdots & h_2^S & l_2^S & d_2^S & & \\
 & & & \ddots & & & & & \ddots & & & & \ddots & \ddots & \ddots & \\
 & & & & k_1^S & & & & & k_2^S & & & & h_N^S & l_N^S & d_N^S
\end{array}\right].$$

Next, we discuss the discretization of the following adjoint equations

$$-\partial_t(p_s w_\alpha) - A_s \partial_x(p_s w_\alpha) = \sum_{j=1}^{S} \mathcal{Q}_{js} p_j w_\alpha, \qquad s \in \mathbb{S}.$$

Replacing $A_s(x, u_s) = a_s x + b_s u_s + c_s$, this set of equations becomes

$$-\partial_t(p_s w_\alpha) = a_s x \partial_x(p_s w_\alpha) + (b_s u_s + c_s)\partial_x(p_s w_\alpha) + \sum_{j=1}^{S} \mathcal{Q}_{js} p_j w_\alpha, \quad s \in \mathbb{S}. \tag{7.14}$$

We denote the approximation of p_s by $p_\Delta^s = \sum_{n=0}^{N} \hat{p}_n^s(t)\tilde{\mathrm{H}}_n(x) \in V_N$. We have

$$\left(p_\Delta^s w_\alpha\right)(x,t) = \sum_{n=0}^{N} \frac{1}{\sqrt{2^n n!}} \hat{p}_n^s(t) H_n(\alpha x). \tag{7.15}$$

We insert (7.15) into (7.14) and obtain

$$-\frac{d}{dt}\hat{p}_n^s(t) = d_n^s \hat{p}_n^s(t) + l_{n+1}^s \hat{p}_{n+1}^s(t) + h_{n+2}^s \hat{p}_{n+2}^s(t) + \sum_{\substack{j=1\\ j\neq s}}^{S} k_s^j \hat{p}_n^j(t), \;\; s \in \mathbb{S}. \tag{7.16}$$

By setting $\hat{p}_{N+1}^s = \hat{p}_{N+2}^s = 0$ for $s \in \mathbb{S}$, equation (7.16), with $0 \le n \le N$, gives us a system of ODEs. We introduce the time transformation $\tau = T - t$, and define

$q_s(t) = p_s(\tau)$ and $\hat{q}_n(t) = \hat{p}_n(\tau)$ to make this system forward in time. The resulting initial-value system can be written in the following matrix form

$$\frac{d\hat{q}_\Delta(t)}{dt} = M_q(T-t)\hat{q}_\Delta(t), \tag{7.17}$$

where

$$\hat{q}_\Delta = [\hat{q}_0^1, \cdots, \hat{q}_N^1, \hat{q}_0^2, \cdots, \hat{q}_N^2, \cdots, \hat{q}_0^S, \cdots, \hat{q}_N^S]^T.$$

The matrix M_q is a $S(N+1) \times S(N+1)$ sparse matrix as follows

$$M_q = \begin{bmatrix}
d_0^1 & l_1^1 & h_2^1 & & & k_1^2 & & & & & & k_1^S & & & & \\
& \ddots & \ddots & \ddots & & & k_1^2 & & & & & & k_1^S & & & \\
& & d_{N-2}^1 & l_{N-1}^1 & h_N^1 & & & k_1^2 & & & \cdots & & & k_1^S & & \\
& & & d_{N-1}^1 & l_N^1 & & & & \ddots & & & & & & \ddots & \\
& & & & d_N^1 & & & & & k_1^2 & & & & & & k_1^S \\
& k_2^1 & & & & d_0^2 & l_1^2 & h_2^2 & & & & k_2^S & & & & \\
& & k_2^1 & & & & \ddots & \ddots & \ddots & & & & k_2^S & & & \\
& & & k_2^1 & & & & d_{N-2}^2 & l_{N-1}^2 & h_N^2 & \cdots & & & k_2^S & & \\
& & & & \ddots & & & & d_{N-1}^2 & l_N^2 & & & & & \ddots & \\
& & & & k_2^1 & & & & & d_N^2 & & & & & & k_2^S \\
& & & \vdots & & & & \vdots & & & \ddots & & & \vdots & & \\
& k_S^1 & & & & k_S^2 & & & & & & d_0^S & l_1^S & h_2^S & & \\
& & k_S^1 & & & & k_S^2 & & & & & & \ddots & \ddots & \ddots & \\
& & & k_S^1 & & & & k_S^2 & & & \cdots & & & d_{N-2}^S & l_{N-1}^S & h_N^S \\
& & & & \ddots & & & & \ddots & & & & & & d_{N-1}^S & l_N^S \\
& & & & k_S^1 & & & & & k_S^2 & & & & & & d_N^S
\end{bmatrix}.$$

The initial data $f_s(x,0) = f_s^0(x)$, $s \in \mathbb{S}$, needs also to be represented in the Hermite functions space by the following

$$\bar{f}_\Delta^s(x) = \sum_{n=0}^{N} \bar{f}_n^s \tilde{\mathrm{H}}_n(x),$$

where

$$\bar{f}_n^s = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} f_s^0(x)\tilde{\mathrm{H}}_n(x)w_\alpha(x)dx, \quad n \geq 0.$$

Since $p_s(x,T) = f_s^T(x) - f_s(x,T)$, after calculating the numerical solution of the forward equations, the initial conditions for the variables q_s, which are the terminal conditions for the adjoint variables p_s, can be approximated as follows

$$\bar{q}_\Delta^s = \sum_{n=0}^{N} \bar{p}_n^s \tilde{\mathrm{H}}_n(x),$$

where

$$\bar{p}_n^s = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} f_s^T(x)\tilde{\mathrm{H}}_n(x)w_\alpha(x)dx - \hat{f}_n^s(T), \quad n \geq 0.$$

Let

$$\bar{f}_\Delta = [\bar{f}_0^1, \cdots, \bar{f}_N^1, \bar{f}_0^2, \cdots, \bar{f}_N^2, \cdots, \bar{f}_0^S, \cdots, \bar{f}_N^S]^T,$$
$$\bar{q}_\Delta = [\bar{q}_0^1, \cdots, \bar{q}_N^1, \bar{q}_0^2, \cdots, \bar{q}_N^2, \cdots, \bar{q}_0^S, \cdots, \bar{q}_N^S]^T.$$

Now the PDE problem of approximating the solution of the forward and backward FP equations in the optimality system (6.12) reduces to the problem of solving the following two linear first-order systems of ODEs,

$$\frac{d\hat{f}_\Delta(t)}{dt} = M_f(t)\hat{f}_\Delta(t), \quad \hat{f}_\Delta(0) = \bar{f}_\Delta,$$

and

$$\frac{d\hat{q}_\Delta(t)}{dt} = M_q(T-t)\hat{q}_\Delta(t), \quad \hat{q}_\Delta(0) = \bar{q}_\Delta.$$

We notice that for the case of constant controls, the solutions $\hat{f}_\Delta$ and $\hat{q}_\Delta$ are given by

$$\hat{f}_\Delta(t) = \exp(M_f t)\,\bar{f}_\Delta \quad \text{and} \quad \hat{q}_\Delta(t) = \exp(M_q t)\,\bar{q}_\Delta.$$

Next, we focus on the control variables u_s. Recall the Hermite expansions

$$f_\Delta^s(x,t) = \sum_{n=0}^{N} \hat{f}_n^s(t)\tilde{\mathrm{H}}_n(x), \quad \text{and} \quad p_\Delta^s(x,t) = \sum_{n=0}^{N} \hat{p}_n^s(t)\tilde{\mathrm{H}}_n(x).$$

We have

$$\begin{aligned}
\int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_\Delta^s) p_\Delta^s w_\alpha dx &= \int_{\mathbb{R}} b_s \partial_x f_\Delta^s p_\Delta^s w_\alpha dx \\
&= b_s \int_{\mathbb{R}} \left(\sum_{n=0}^{N} \hat{f}_n^s \partial_x \tilde{\mathrm{H}}_n(x)\right)\left(\sum_{n=0}^{N} \hat{p}_n^s \tilde{\mathrm{H}}_n(x)\right) w_\alpha(x) dx \\
&= -b_s\alpha \int_{\mathbb{R}} \left(\sum_{n=0}^{N} \sqrt{2(n+1)}\hat{f}_n^s \tilde{\mathrm{H}}_{n+1}(x)\right)\left(\sum_{n=0}^{N} \hat{p}_n^s \tilde{\mathrm{H}}_n(x)\right) w_\alpha(x) dx \\
&= -b_s\alpha \sum_{n=0}^{N}\sum_{k=0}^{N} \sqrt{2(n+1)}\hat{f}_n^s \hat{p}_k^s \int_{\mathbb{R}} \tilde{\mathrm{H}}_{n+1}(x)\tilde{\mathrm{H}}_k(x) w_\alpha(x) dx \\
&= -b_s\alpha \sum_{n=0}^{N-1} \sqrt{2(n+1)}\hat{f}_n^s \hat{p}_{n+1}^s \frac{\sqrt{\pi}}{\alpha} \\
&= -b_s\sqrt{2\pi} \sum_{n=0}^{N-1} \sqrt{n+1}\hat{f}_n^s \hat{p}_{n+1}^s.
\end{aligned}$$

Therefore, the optimality condition $\nu u_s + \int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_s) p_s w_\alpha \, dx = 0$ gives the following

$$\begin{aligned}
u_\Delta^s(t) &= -\frac{1}{\nu}\int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_\Delta^s) p_\Delta^s w_\alpha dx \\
&= \frac{b_s\sqrt{2\pi}}{\nu} \sum_{n=0}^{N-1} \sqrt{n+1}\hat{f}_n^s \hat{p}_{n+1}^s.
\end{aligned}$$

For constant controls, we have

$$\begin{aligned} u_\Delta^s &= -\frac{1}{\nu}\int_0^T \int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_\Delta^s) p_\Delta^s w_\alpha dx\, dt \\ &= \frac{b_s\sqrt{2\pi}}{\nu}\sum_{n=0}^{N-1}\sqrt{n+1}\int_0^T \hat{f}_n^s \hat{p}_{n+1}^s\, dt. \end{aligned}$$

7.2.1 Convergence analysis

In this section, we discuss the approximation properties of our Hermite discretization scheme applied to the FP optimality system (7.9). To analyze the accuracy of the scheme for constant controls, we first show that the approximations for the state and adjoint variables are spectrally convergent.

Theorem 18. *If $f_s, p_s \in L^\infty(0,T; H^r_{w_\alpha}(\mathbb{R}))$ for $s \in \mathbb{S}$ and $r > 1$, and the control variables are constant, then for all $t \in [0,T]$ the following holds*

$$\|f_s(\cdot,t) - f_\Delta^s(\cdot,t)\|_{w_\alpha}^2 = O(N^{-r}) \quad \textit{and} \quad \|p_s(\cdot,t) - p_\Delta^s(\cdot,t)\|_{w_\alpha}^2 = O(N^{-r}).$$

Proof. For any $s \in \mathbb{S}$, we have

$$\begin{aligned} f_s(x,t) - f_\Delta^s(x,t) &= \sum_{n=0}^{\infty} \hat{f}_n^s(t)\tilde{\mathrm{H}}_n(x) - \sum_{n=0}^{N} \hat{f}_n^s(t)\tilde{\mathrm{H}}_n(x) \\ &= \sum_{n=N+1}^{\infty} \hat{f}_n^s(t)\tilde{\mathrm{H}}_n(x). \end{aligned}$$

Notice that for constant controls no truncation error appears in calculating the Hermite coefficients $\hat{f}_n^s$ by solving the ODE system (7.13). This is because of the fact that for $m > n$ the value of $\hat{f}_n^s$ is independent of the value of $\hat{f}_m^s$. That is, $P_N f_s(\cdot,t) = f_\Delta^s(\cdot,t)$ for every $t \in [0,T]$. So we have the following bound for the error

$$\begin{aligned} \|f_s(\cdot,t) - f_\Delta^s(\cdot,t)\|_{w_\alpha}^2 &= \int_{\mathbb{R}} \left[\sum_{n=N+1}^{\infty} \hat{f}_n^s(t)\tilde{\mathrm{H}}_n(x)\right]^2 w_\alpha(x) dx \\ &= \frac{\sqrt{\pi}}{\alpha}\sum_{n=N+1}^{\infty} |\hat{f}_n^s(t)|^2 \\ &\le \frac{\sqrt{\pi}}{\alpha}\sum_{n=N+1}^{\infty} \frac{\alpha^{2-2r}}{\pi} n^{-r} \|f_s(\cdot,t)\|_{r,w_\alpha}^2. \end{aligned}$$

Therefore, we have $\|f_s(\cdot,t) - f_\Delta^s(\cdot,t)\|_{w_\alpha}^2 = O(N^{-r})$.

Regarding the adjoint variables p_s, $s \in \mathbb{S}$, we note that the coefficient matrix M_q is the transpose of the matrix M_f. Therefore, in the framework of discretize-before-optimize [20], we have the same estimate for the adjoint variables. □

Next, we investigate the spectral convergence in approximating the constant control variables in our bilinear control mechanism.

Theorem 19. *Let $f_s \in L^\infty(0,T; H^r_{w_\alpha}(\mathbb{R}))$, $s \in \mathbb{S}$, $r > 2$ and $N \geq 2$. Then for a positive constant c_s, we have the estimate*

$$|u_s - u^s_\Delta| \leq c_s \sum_{n=N}^{\infty} n^{1-r}.$$

Proof. Consider the optimality equations in the continuous and discrete form. We have

$$\begin{aligned} \nu u_s + \int_0^T \int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_s)\, p_s\, w_\alpha\, dx\, dt = 0, \\ \nu u^s_\Delta + \int_0^T \int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f^s_\Delta)\, p^s_\Delta\, w_\alpha\, dx\, dt = 0. \end{aligned}$$

We note that

$$\begin{aligned} \int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_s)\, p_s\, w_\alpha dx &= \int_{\mathbb{R}} b_s \left(\sum_{n=0}^{\infty} \hat{f}^s_n \partial_x \tilde{\mathrm{H}}_n(x) \right) \left(\sum_{n=0}^{\infty} \hat{p}^s_n \tilde{\mathrm{H}}_n(x) \right) w_\alpha dx \\ &= -b_s \alpha \int_{\mathbb{R}} \left(\sum_{n=0}^{\infty} \sqrt{2(n+1)} \hat{f}^s_n \tilde{\mathrm{H}}_{n+1}(x) \right) \left(\sum_{n=0}^{\infty} \hat{p}^s_n \tilde{\mathrm{H}}_n(x) \right) w_\alpha dx \\ &= -b_s \alpha \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \sqrt{2(n+1)} \hat{f}^s_n \hat{p}^s_k \int_{\mathbb{R}} \tilde{\mathrm{H}}_{n+1}(x) \tilde{\mathrm{H}}_k(x) w_\alpha dx \\ &= -b_s \alpha \sum_{n=0}^{\infty} \sqrt{2(n+1)} \hat{f}^s_n \hat{p}^s_{n+1} \frac{\sqrt{\pi}}{\alpha} \\ &= -b_s \sum_{n=0}^{\infty} \sqrt{2\pi(n+1)} \hat{f}^s_n \hat{p}^s_{n+1}. \end{aligned}$$

Similarly, we have

$$\int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f^s_\Delta)\, p^s_\Delta w_\alpha dx = -b_s \sum_{n=0}^{N-1} \sqrt{2\pi(n+1)} \hat{f}^s_n \hat{p}^s_{n+1}.$$

Therefore, we obtain

$$\begin{aligned} &- \int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_s) p_s w_\alpha dx + \int_R \partial_x(\partial_{u_s} A_s f^s_\Delta)\, p^s_\Delta w_\alpha dx \\ &= b_s \sum_{n=0}^{\infty} \sqrt{2\pi(n+1)} \hat{f}^s_n \hat{p}^s_{n+1} - b_s \sum_{n=0}^{N-1} \sqrt{2\pi(n+1)} \hat{f}^s_n \hat{p}^s_{n+1} \\ &= b_s \sum_{n=N}^{\infty} \sqrt{2\pi(n+1)} \hat{f}^s_n \hat{p}^s_{n+1}. \end{aligned}$$

Hence

$$|\int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f_s)\, p_s w_\alpha dx - \int_{\mathbb{R}} \partial_x(\partial_{u_s} A_s f^s_\Delta)\, p^s_\Delta w_\alpha dx| \le |b_s|\sqrt{2\pi} \sum_{n=N}^{\infty} \sqrt{n+1}|\hat{f}^s_n|.|\hat{p}^s_{n+1}|.$$

Assuming $N \ge 2$, Lemma 7 gives us the following

$$\begin{aligned}\sum_{n=N}^{\infty} \sqrt{n+1}|\hat{f}^s_n(t)|.|\hat{p}^s_{n+1}(t)| &\le \sum_{n=N}^{\infty} \sqrt{n+1}(\frac{\alpha^{1-r}}{\sqrt{\pi}})^2 n^{-r/2}(n+1)^{-r/2}\|f_s(\cdot,t)\|_{r,w_\alpha}\|p_s(\cdot,t)\|_{r,w_\alpha} \\ &\le c^s_{fp} \sum_{n=N}^{\infty} n\, n^{-r/2} n^{-r/2} = c^s_{fp} \sum_{n=N}^{\infty} n^{1-r},\end{aligned}$$

in which $c^s_{fp} = \frac{\alpha^{2-2r}}{\pi}\|f_s(\cdot,t)\|_{r,w_\alpha}\|p_s(\cdot,t)\|_{r,w_\alpha}$. Therefore, the desired accuracy estimate for the control variable u_s can be written as follows

$$|u_s - u^s_\Delta| \le c^s \sum_{n=N}^{\infty} n^{1-r},$$

where $c^s = |b_s|\frac{\sqrt{2\pi}}{\nu}\int_0^T c^s_{fp}\, dt.$ □

7.2.2 Numerical experiments

We present results of numerical experiments to validate the accuracy of the Hermite-spectral discretization for approximating the solution of the PDP FP optimality system. We anticipate that the results of numerical implementation depend on the value assigned to the scaling factor α introduced in the weight function $\exp(\alpha^2 x^2)$ which also appears in $H_n(\alpha x)$ to produce the n-th Hermite function. It is still an open problem how to find the optimal scaling factor, in spite of significant attentions payed to the importance of this issue in scientific literature; see, e.g., [45, 74, 75]. When dealing with a Gaussian function as an initial condition or as the equilibrium solution, it is possible to follow the arguments presented in [45] and [74] to select a suitable scaling factor. Concerning a Gaussian function like $\exp(-\beta x^2)$, we can set $\alpha = \sqrt{\beta}$ which is a simple choice satisfying the necessary condition $\alpha < \sqrt{2\beta}$ prescribed in the above-mentioned references. However, concerning time dependent problems, where the characteristics of the solution change during the time evolution, the best scaling factor for representing the initial solution may depend on time. It means that for evolutionary problems it may be advantageous to consider a time dependent scaling factor, which is beyond the scope of this work. We find a proper scaling factor by experimental trials.

To examine our discretization scheme for the optimal control problem we insert time-dependent controls u_1 and u_2 into the FP equations (5.24)-(5.25). We specifically have $A_1(x,u_1) = -x+u_1+1$ and $A_2(x,u_2) = -x-u_2-1$, and introduce the following target functions

$$f_1^T(x) = \frac{1}{2\sqrt{2\pi\sigma^2}}\exp(\frac{-(x-1+e^{-T})^2}{2\sigma^2}),$$

and

$$f_2^T(x) = \frac{1}{2\sqrt{2\pi\sigma^2}} \exp(\frac{-(x+1-e^{-T})^2}{2\sigma^2}).$$

We set $\nu = 0.1$, $\sigma^2 = 0.05$, $\mu = 2$, $\gamma = 1$, $W = 1$, $T = 10$, $\alpha = 1.5$, and aim to approach as close as possible to the target functions f_1^T and f_2^T starting with the initial conditions

$$f_1^0(x) = f_2^0(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(\frac{-x^2}{2\sigma_0^2}),$$

with $\sigma_0^2 = 0.1$. To solve the optimality system (6.12) with this setting and time-dependent controls, we apply the nonlinear conjugate gradient method proposed in [7, 20]. The results of approximating the state variables f_1 and f_2 are depicted in Figures 7.2 and 7.3, which illustrate how the solutions improve by increasing the number of expansion terms.

Figure 7.4 refers to the approximation of the adjoint variables p_1 and p_2 corresponding to $N = 50$. In Figure 7.5, we can follow the time evolution of the control variables u_1 and u_2. Since both f_1 and f_2 have the same initial PDFs and also follow the same Gaussian targets which are centered differently, the controls u_1 and u_2 coincide. Similar results are obtained in the case of different targets; see Figures 7.6-7.7. In this experiment, we have considered the same setting as the previous test, with the only difference that $\sigma^2 = 0.05$ in f_1^T while $\sigma^2 = 0.1$ in f_2^T.

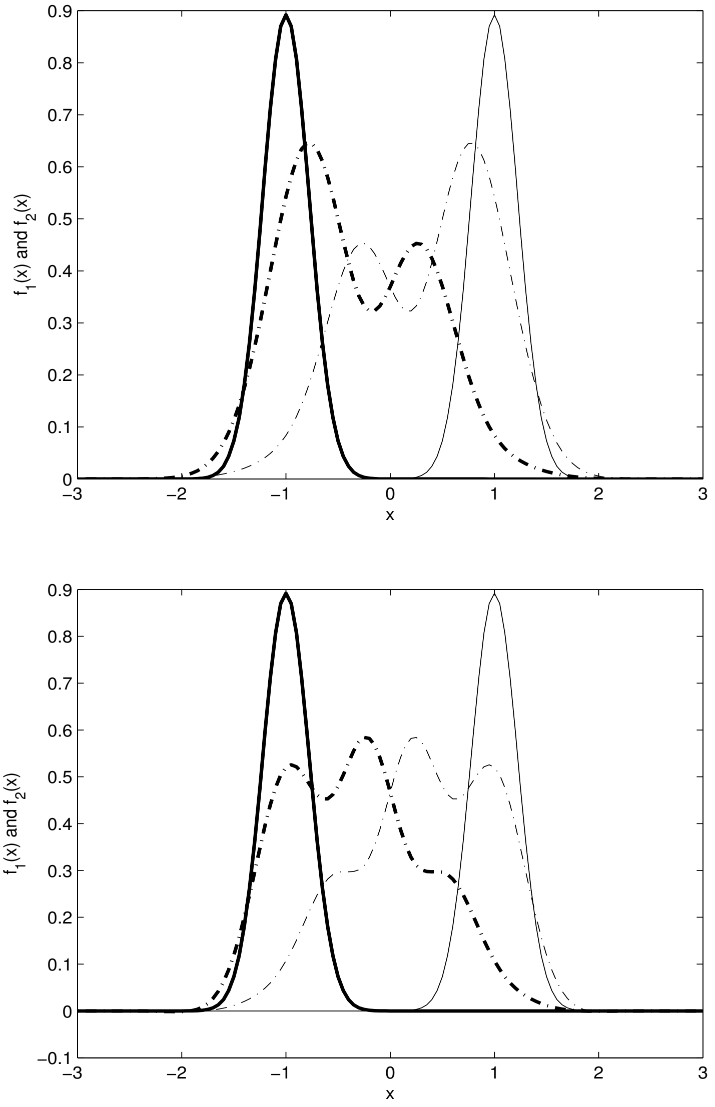


Figure 7.2: Numerical solutions to the states of the optimality system (dash-dot lines) and the target functions (solid lines) at time $t = T$; top: $N = 5$, bottom: $N = 10$; $\alpha = 1.5$. Thin lines correspond to f_1 and thick lines correspond to f_2.

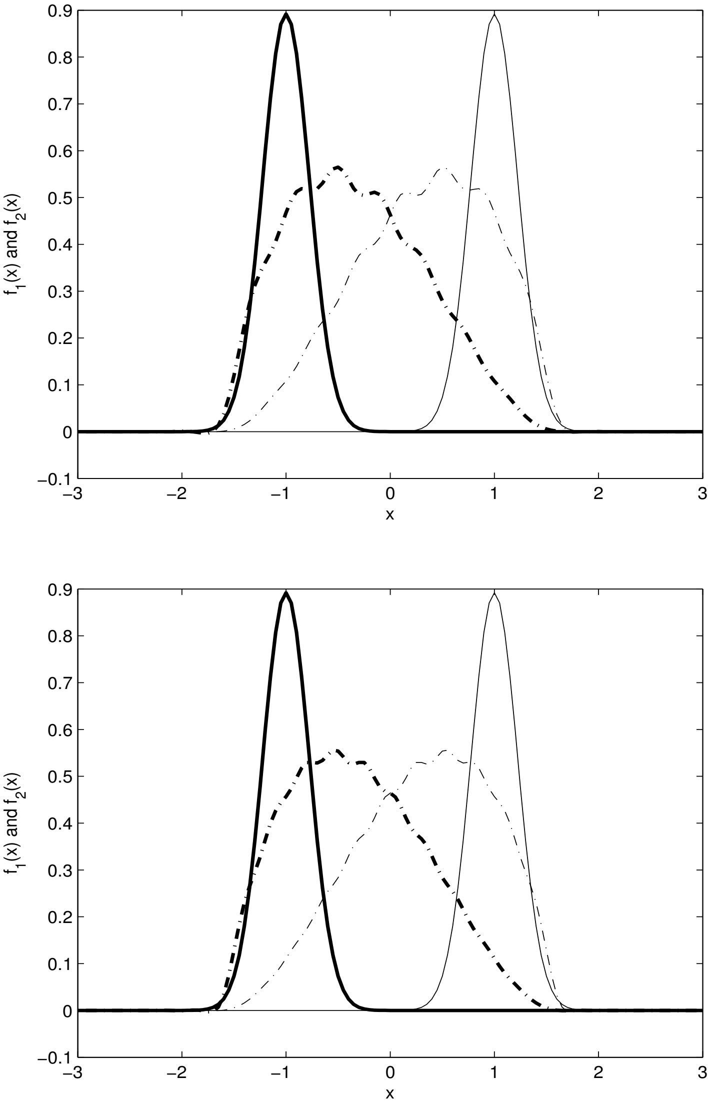
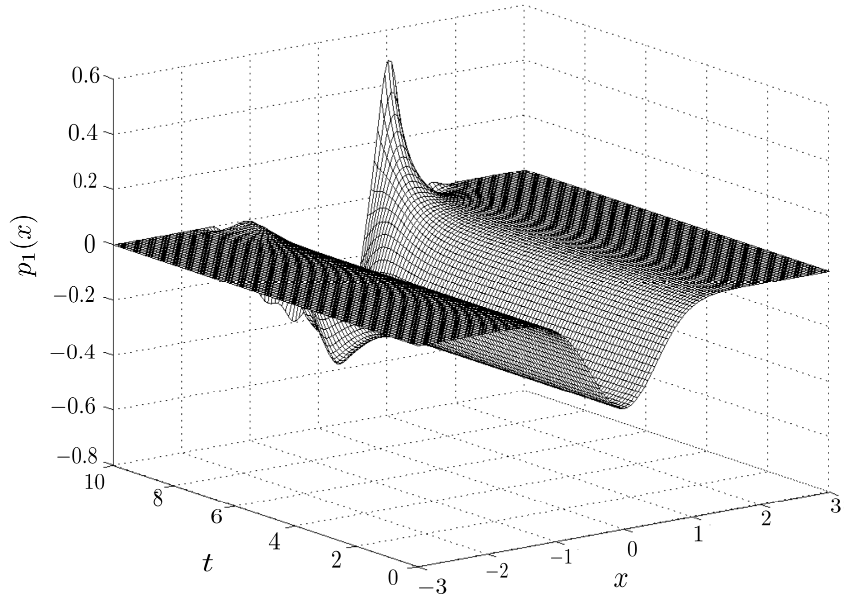


Figure 7.3: Numerical solutions to the states of the optimality system (dash-dot lines) and the target functions (solid lines) at time $t = T$; top: $N = 50$, bottom: $N = 100$; $\alpha = 1.5$. Thin lines correspond to f_1 and thick lines correspond to f_2.



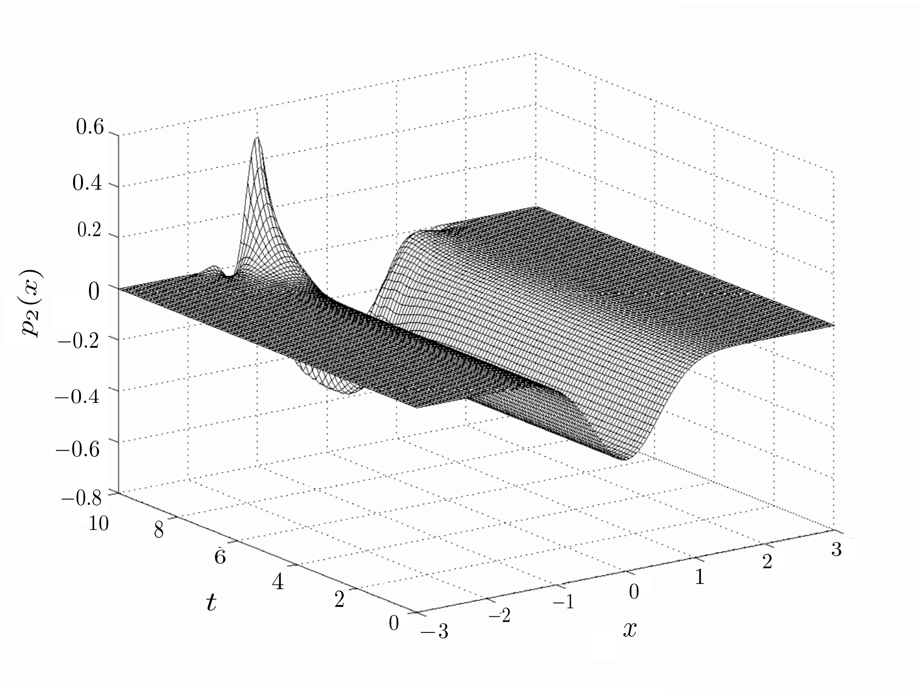


Figure 7.4: Numerical solutions to the adjoint variables p_1 (top) and p_2 (bottom); $N = 50$, $\alpha = 1.5$.

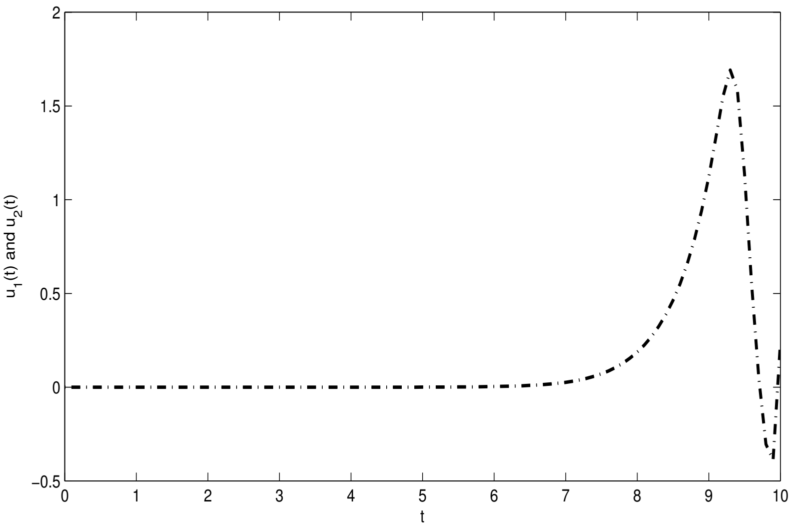


Figure 7.5: Numerical solutions to the time-dependent control variables; $N = 50$, $\alpha = 1.5$. Thin line corresponds to u_1 and thick line corresponds to u_2 (both solutions coincide).

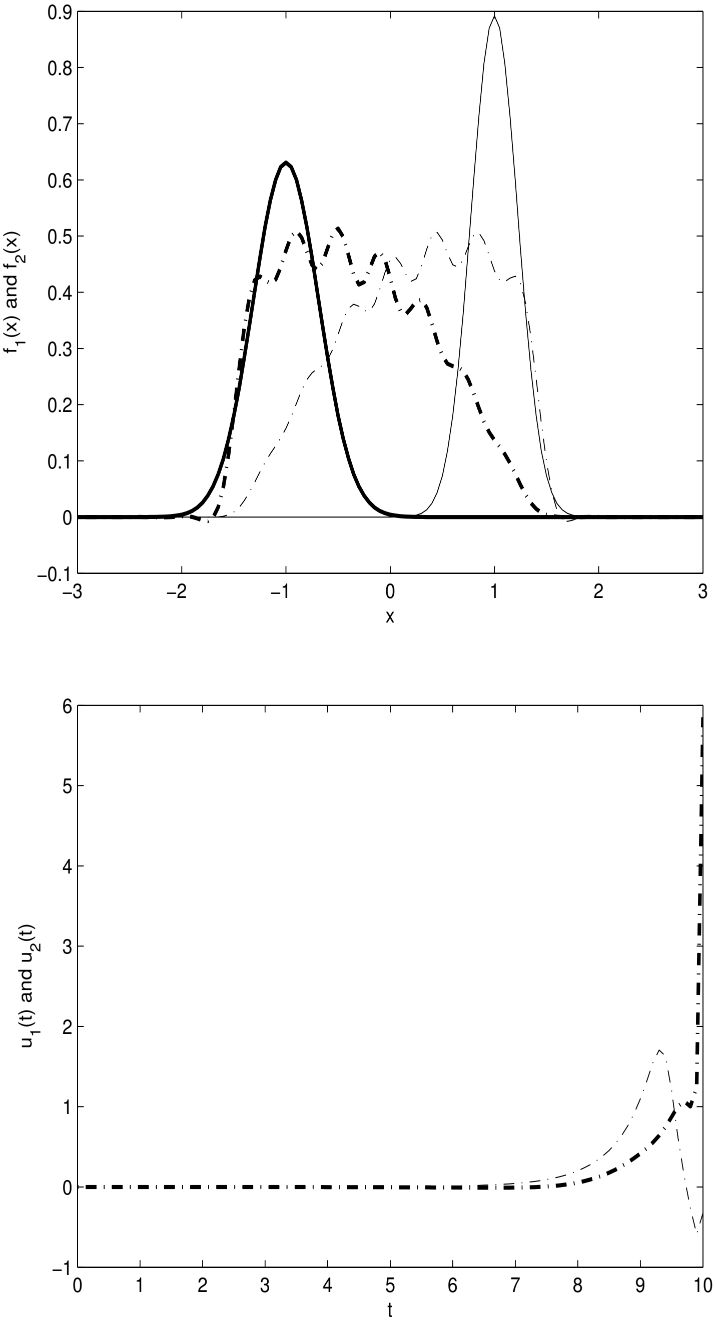
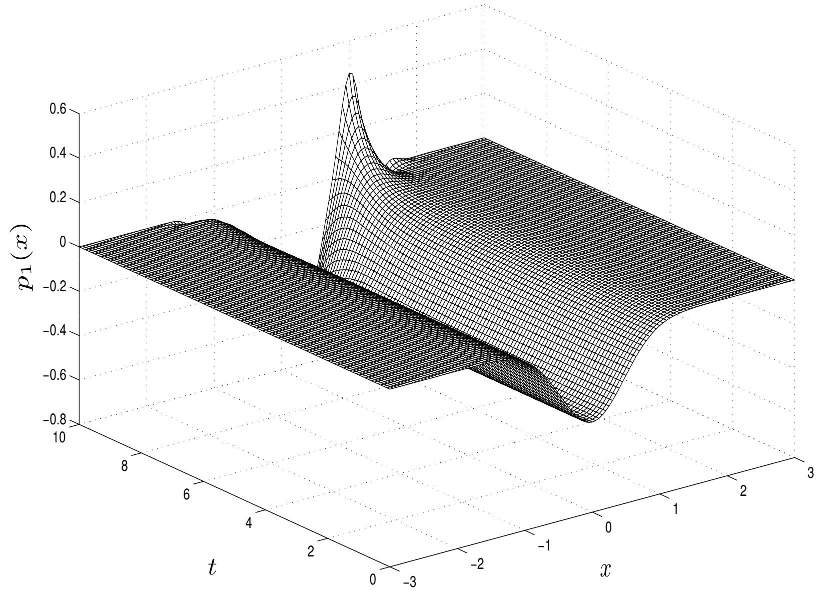


Figure 7.6: Top: Numerical solutions to the states of the optimality system (dash-dot lines) and the target functions (solid lines) at time $t = T$. Bottom: Numerical solutions to the time-dependent control variables. Thin lines correspond to f_1 and u_1 and thick lines correspond to f_2 and u_2; $N = 50$, $\alpha = 1.5$.



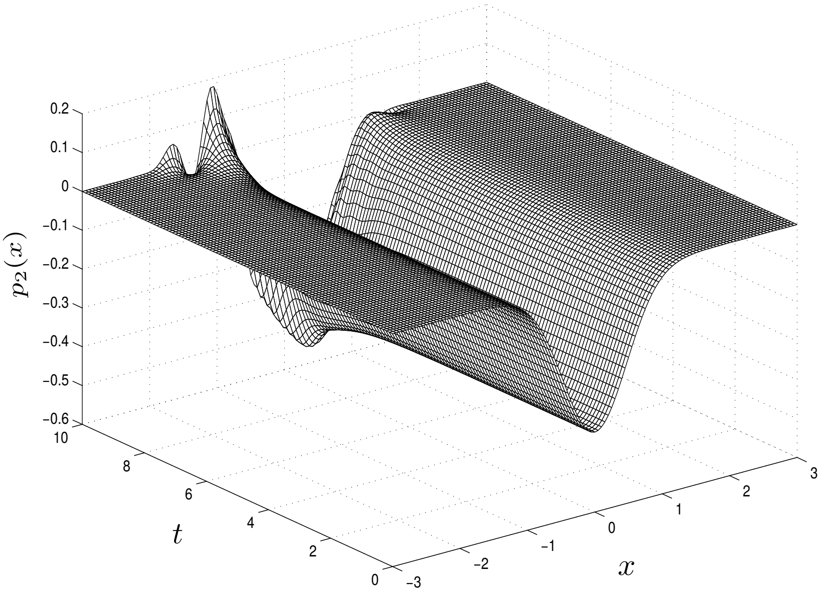


Figure 7.7: Numerical solutions to the adjoint variables p_1 (top) and p_2 (bottom) when different targets are prescribed; $N = 50$, $\alpha = 1.5$.

Chapter 8

Conclusion

The Fokker-Planck equations which we considered in this thesis are mathematical models describing the time evolution of the probability density function of Itō stochastic processes and piecewise deterministic processes. We therefore summarized at the beginning of the thesis the most important concepts which were necessary to have an intuitive understanding of these stochastic processes. We then discussed the FP equations as second order partial differential equations. In particular, we illustrated how to derive the FP equations of parabolic and hyperbolic type from corresponding stochastic process.

After establishing the fundamental definitions and setting for FP equations, we discretized the equations in both bounded and unbounded domains. In the case of a bounded domain, we considered and discussed finite difference discretization schemes to approximate the solution of FP equations. Particularly, for parabolic FP equations, finite difference schemes based on the Chang-Cooper method and first- and second-order backward time differencing schemes were investigated. These schemes provide conditionally stable solutions that are second-order accurate in space and first- and second-order accurate in time, respectively. Moreover, these schemes satisfy conservation and positivity properties of the Fokker-Planck solution. These properties were theoretically proven and validated by numerical experiments.

Next, we investigated the Hermite spectral discretization of the FP equations, both parabolic and hyperbolic types, defined on unbounded domains. First the required properties and equipment for spectral methods and particularly Hermite approximation were discussed. Then the discretization schemes were analyzed, and the accuracy of the Hermite spectral method was proved by showing that, in the both cases, the error decreases spectrally as the number of expansion terms increases. Moreover, it was proved that the proposed discretization schemes preserve conservativity of the FP solutions. Results of numerical experiments demonstrated the theoretical estimates. Since a weighted Hermite approximation method was used, the optimal choice for the scaling factor in the weight function was investigated with numerical experiments.

The rest of the thesis was dedicated to the optimal control problems related to the FP equations. We first shortly introduced the optimal control problems, and discussed the optimal control problems governed by FP equations to find controls with the purpose of driving the random processes to attain desired objectives. For the both cases, Itō stochastic models and PDP processes, we derived the corresponding FP optimality systems consisting of the state, the adjoint, and the optimality condition equations. We then investigated the Hermite spectral discretization of these FP optimality systems in unbounded domains. The accuracy of the discretization scheme was discussed by showing spectral convergence in approximating the state, the adjoint, and the control variables that appear in the FP optimality systems. To further investigate the effectiveness of the method, results of numerical experiments were presented as well.

Chapter 9

Appendix. A model predictive control scheme

Our purpose is to define a control strategy for the probability density function of a stochastic process to track a given sequence of desired PDFs in time. In mathematical terms, this means to minimize the tracking objective at a given time instants. Let $(0, T)$ be the time interval where the process is considered. We assume time windows of size $\Delta t = T/N$ with N a positive integer. Let $t_k = k\Delta t$, $k = 0, 1, \ldots, N$. At time t_0, we have a given initial PDF denoted with ρ and with $f_d(\cdot, t_k)$, $k = 1, \ldots, N$, we denote the sequence of desired PDFs. Our scheme starts at time t_0 and solves the minimization problem $\min_u J(f(u), u)$ defined in the interval (t_0, t_1). Then, with the probability density function f resulting at $t = t_1$ that solves the optimal control problem in (t_0, t_1), we define the initial PDF for the subsequent optimization problem defined in the interval (t_1, t_2). This procedure is repeated by receding the time horizon until the last time window is reached. This is an instance of the class of receding horizon model predictive control (RH-MPC) schemes [76, 79] that is widely used in engineering applications to design closed-loop algorithms. In fact, we implement a MPC scheme where the time horizon used to evaluate the control coincides with the time horizon where the control is used. One important aspect of this approach is that it can be applied to infinite dimensional evolution systems [59], that is the case of the FP model. We refer to [86] to show that the closed-loop system with the RH-MPC scheme is nominally asymptotically stable.

The RH-MPC procedure is summarized in the following algorithm.

Algorithm 20 (RH-MPC Control)**.** *Set $k = 0$, $\rho_0 = \rho$;*

1. *Assign the initial PDF, $f(x, t_k) = \rho_k(x)$ and the target $f_d(\cdot, t_{k+1})$;*
2. *In (t_k, t_{k+1}), apply Algorithm 22 to solve $\min_{u\in\mathbb{R}^\ell} J(f(u), u)$, thus obtain the optimal pair (f, u);*
3. *If $t_{k+1} < T$, set $k := k + 1$, $\rho_k = f(\cdot, t_k)$, go to 1.*
4. *End.*

Next, we discuss the first step of Algorithm 20, that consists in solving $\min_u J(f(u), u)$. In fact, the solution of the FP control problem given by the mapping $u \to y(u)$ allows to transform the constrained optimization problem in an unconstrained one as follows

$$\min_{u \in U} \hat{J}(u). \tag{9.1}$$

Thus the solution to the FP equation is included into the objective at $u \in U \subset \mathbb{R}^\ell$. Further, to compute $\nabla_u \hat{J}(u)$ for a given u, we have to solve first the forward FP equation and then the adjoint FP equation. This procedure is summarized in the following

Algorithm 21 (Evaluation of the gradient at u)**.**

1. *Solve the discrete FP equation with given initial condition;*
2. *Solve the discrete adjoint FP equation with given terminal condition;*
3. *Compute the gradient* $\nabla_u \hat{J}(u)$ *using (6.7);*
4. *End.*

It is clear that the solution of the FP optimality system may become prohibitive when high-dimensional stochastic processes are considered. In this case, special techniques for solving high-dimensional partial differential equations are in order; see, e.g., [53, 110].

We solve the optimization problem by implementing the gradient given by Algorithm 21 in a nonlinear conjugate gradient (NCG) scheme. Nonlinear conjugate gradient schemes represent extensions of linear conjugate gradient methods to non-quadratic problems; see, e.g., [48, 94]. In the common variants, the basic idea is to avoid matrix operations and express the search directions recursively as

$$d_{k+1} = -g_{k+1} + \beta_k \, d_k, \tag{9.2}$$

where $g_k = \nabla \hat{J}(u_k)$, $k = 0, 1, 2, \ldots$, with $d_0 = -g_0$. The iterates for a minimum point are given by

$$u_{k+1} = u_k + \alpha_k \, d_k, \tag{9.3}$$

where $\alpha_k > 0$ is a steplength that satisfies the Armijo condition of sufficient decrease of $\hat{J}$'s value as follows

$$\hat{J}(u_k + \alpha_k \, d_k) \leq \hat{J}(u_k) + \delta \, \alpha_k \, (\nabla \hat{J}(u_k), d_k)_U \tag{9.4}$$

where $0 < \delta < 1/2$; see [84]. Notice that we use the inner product of the $U = \mathbb{R}^\ell$ space.

The parameter β_k is chosen so that (9.2)–(9.3) reduces to the linear CG scheme if $\hat{J}$ is a strictly convex quadratic function and α_k is the exact one-dimensional minimizer of $\hat{J}$ along d_k. In this case the NCG scheme terminates in at most n steps in exact arithmetic. This case provides a lower bound to the computational complexity of NCG schemes.

There are many different formula for β_k which result in different performance depending on the (nonlinear) problem. We use the formulation due to Dai and Yuan [35] as follows

$$\beta_k^{DY} = \frac{(g_{k+1}, g_{k+1})_U}{(d_k, y_k)_U}, \tag{9.5}$$

where $y_k = g_{k+1} - g_k$.

The NCG scheme is implemented as follows.

Algorithm 22 (NCG Scheme)**.**

```
• Input: initial approx. $u_0$, $d_0 = -\nabla \hat{J}(u_0)$, index $k = 0$, maximum $k_{max}$, tolerance $tol$.

    1. While ($k < k_{max}$ && $\|g_k\|_{\mathbb{R}^\ell} > tol$ ) do
    2. Search steplength $\alpha_k > 0$ along $d_k$ satisfying (9.4);
    3. Set $u_{k+1} = u_k + \alpha_k\, d_k$;
    4. Compute $g_{k+1} = \nabla \hat{J}(u_{k+1})$ using Algorithm 21;
    5. Compute $\beta_k^{DY}$ given by (9.5);
    6. Let $d_{k+1} = -g_{k+1} + \beta_k^{DY}\, d_k$;
    7. Set $k = k + 1$;
    8. End while
```

Bibliography

[1] Y. Abe, S. Ayik, P.-G. Reinhard, and E. Suraud, *On stochastic approaches of nuclear dynamics*, Physics Reports, **275** (1996), 49-196.

[2] A. Amudevar, *A dynamic programming algorithm for the optimal control of piecewise deterministic Markov processes*, SIAM J. Control Optim. **40** (2001), 525-539.

[3] M. Annunziato, *On the action of a semi-Markov process on a system of ordinary differential equations*, Math. Mod. Anal., **17** (2012), 650-672.

[4] M. Annunziato, *Analysis of upwind method for piecewise deterministic Markov processes*, Comp. Meth. Appl. Math., **8** (2008), 3-20.

[5] M. Annunziato, *Non-gaussian equilibrium distributions arising from the Langevin equation*, Phys. Rev. E, **65** 21113 (2002), 1-6.

[6] M. Annunziato and A. Borzì, *Optimal control of probability density functions of stochastic processes.* Mathematical Modelling and Analysis, **15** (2010), 393-407.

[7] M. Annunziato and A. Borzì, *A Fokker-Planck control framework for multidimensional stochastic processes.* Journal of Computational and Applied Mathematics, **237** (2013), 487-507.

[8] M. Annunziato and A. Borzì, *Optimal control of a class of piecewise deterministic processes*, European Journal of Applied Mathematics, **25** (2014), 1-25.

[9] M. Annunziato, P. Grigolini and B. J. West, *Canonical and noncanonical equilibrium distribution*, Phys. Rev. E, **66** 011107 (2001), 1-13.

[10] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, John Wiley & Sons, 1974.

[11] D. G. Aronson, *Non-negative solutions of linear parabolic equations.* Annali della Scuola Normale Superiore di Pisa, Classe di Scienze 3^a serie (1968) **22**, 607-694.

[12] N. Bäurle and U. Rieder, *MDP Algorithms for portfolio optimization problems in pure jump markets*, Finance and Stochastics, **13** (2009), 591-611.

[13] R. Blackmore and B. Shizgal, *Discrete-ordinate method of solution of Fokker-Planck equations with nonlinear coefficients*, Phys. Rev. A, **31** (1985), 1855.

[14] R. Blackmorea, U. Weinerta and B. Shizgal, *Discrete ordinate solution of a Fokker-Planck equation in laser physics*, Transport Theory and Statistical Physics, **15** (1986), 181-210.

[15] T. Blum and A. J. McKane, *Variational schemes in the Fokker-Planck equation*, Journal of Physics A: Mathematical and General, **29** (1996), 1859.

[16] V. Bogachev, G. Da Prato, M. Röckner, *Existence and uniqueness of solutions for Fokker-Planck equations on Hilbert spaces*, J. Evol. Equations, **10** (2010) 487-509.

[17] R. Bonifacio and L. A. Lugiato, *Dissipative Systems in Quantum Optics*, Springer, Berlin, (1982), 329-360.

[18] C.L. Bris, P.L. Lions, *Existence and uniqueness of solutions to Fokker-Planck type equations with irregular coefficients.* Communications in Partial Differential Equations (2008) **33**, 1272-1317.

[19] C. Bolley, M. Crouzeix, *Conservation de la positivite lors de la discretisation des problemes d'evolution paraboliques.* Analyse numerique, **12** (1978), 237-245.

[20] A. Borzi and V. Schulz, *Computational optimization of systems governed by partial differential equations*, SIAM book series on Computational Science and Engineering 08, SIAM, Philadelphia, PA, 2012.

[21] C. Buet, and S. Cordier, and P. Degond, and M. Lemou, *Fast algorithms for numerical, conservative, and entropy approximations of the Fokker-Planck-Landau equation.* J. Comput. Phys. (1997) **133**, 310-322.

[22] C.G. Cassandras and J. Lygeros, *Stochastic Hybrid Systems* (CRC Press Taylor & Francis group, 2007).

[23] J. S. Chang and G. Cooper, *A Practical Difference Scheme for Fokker-Planck Equation.* Journal of Computational Physics, **6** (1970), 1-16.

[24] T. Chen, *A theoretical and numerical study for the Fokker-Planck equation*, Master thesis, Simon Fraser university, 1992.

[25] K. H. Chew, P. N. Shivakumar, and J. J. Williams, *Error Bounds for the Truncation of Infinite Linear Differential Systems*, J. Inst. Maths Applics, **25** (1980) 37-51.

[26] D. R. Chialvo and A. V. Apkarian, *Modulated noisy biological dynamics: three examples*, J. Stat. Phys., **70** (1993), 375-391.

[27] J. Chiquet, N. Limnios and M. Eid, *Piecewise deterministic markov processes applied to fatigue crack growth modelling*, Journal of Statistical Planning and Inference, (Special Issue on Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing), **139** (2009), 1657-1667.

[28] K. G. Choo, K. L. Teo and Z. S. Wu, *On an optimal control problem involving first order hyperbolic systems with boundary controls*, Numer. Funct. anal. Optim., **4** (1982), 171-190.

[29] C. Cocozza-Thivent, R. Eymard, S. Mercier and M. Roussignol, *Characterization of the marginal distributions of Markov processes used in dynamic reliability*, Journal of Applied Mathematics and Stochastic Analysis, (2006), 1-18.

[30] W. T. Coffey and D. S. F. Crothers, *Comparison of methods for the calculation of superparamagnetic relaxation times*, Phys. Rev. E, **54** (1996), 4768.

[31] O. L. V. Costa and F. Dufour, *Average continuous control of piecewise deterministic Markov processes*, SIAM J. Control Optim, **48** (2010), 4262-4291.

[32] O. L.V. Costa and F. Dufour, *On the Poisson equation for piecewise-deterministic Markov processes*, SIAM J. Control Optim., **42** (2003) 985-1001.

[33] D.R. Cox and H.D. Miller, *The Theory of Stochastic Processes* Chapman & Hall CRC, 2001.

[34] H.I. M. Crandall, P.L. Lions, *Users guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., **27** (1992), 1-67.

[35] Y.H. Dai and Y. Yuan, *A nonlinear conjugate gradient with a strong global convergence property*, SIAM J. Opt., 10 (1999), 177-182.

[36] M. H. A. Davis, *Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models*, J. R. Stat. Soc., **46** (1984), 353-388.

[37] J. D. Densmore, and J. S. Warsa, and R.B. Lowrie, and J. E. Morel, *Stability analysis of implicit time discretizations for the Compton-scattering Fokker-Planck equation.* Journal of Computational Physics, **228** (2009), 5933-5960.

[38] W. Ebeling, E. Gudowska-Nowak, and I.M. Sokolov, *On stochastic dynamics in physics - Remarks on history and terminology*, Acta Physica Polonica B, **39** (2008), 1003-1018.

[39] E. Emmrich, *Two-step BDF time discretisation of nonlinear evolution problems governed by monotone operators with strongly continuous perturbations.* Computational Methods in Applied Mathematics, **9** (2009), 37-62.

[40] J. C. Englund, W. C. Schieve, W. Zurek, R. F. Gragg , *Fluctuations and Transitions in the Absorptive Optical Bistability*, Optical Bistability, (1981), 315-335.

[41] L.C. Evans, *Partial Differential Equations, Graduate Studies in Mathematics, vol. 19*, American Mathematical Society, Providence, Rhode Island, 2002.

[42] A. Faggionato, D. Gabrielli and M. Ribezzi Crivellari, *Non-equilibrium thermodynamics of piecewise deterministic Markov processes*, Journal of Statistical Physics, **137** (2009), 259-304.

[43] R. Filliger and M. O. Hongler, *Supersymmetry in random two-velocity processes*, Physica. A, **332** (2004), 141-150.

[44] W. Fleming, H. Soner, *Controlled Markov Processes and Viscosity Solutions.* Springer-Verlag 2006.

[45] J. C. M. Fok, B. Guo and T. Tang, *Combined Hermite spectral-finite difference method for the Fokker-Planck equation*, Mathematics of Computation, **71** (2001), 1497-1528.

[46] D. Funaro and O. Kavian, *Approximation of some diffusion evolution equations in unbounded domains by Hermite functions*, Mathematics of Computation, **57** (1991), 597-619.

[47] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Springer-Verlag Berlin Heidelberg, 1983.

[48] J.C. Gilbert and J. Nocedal, *Global convergence properties of conjugate gradient methods for optimization*, SIAM J. Opt., 2 (1992), 21-42.

[49] D. Gottlieb, S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, Society for Industrial and Applied Mathematics, 1977.

[50] M. Grote and J. Keller, *On nonreflecting boundary conditions*, J Comput Phys, **122** (1995), 231-243.

[51] L. GrÃ$\frac{1}{4}$ne and J. Pannek, *Nonlinear Model Predictive Control - Theory and Algorithms*, Springer-Verlag, London, 2011.

[52] B-Y. Guo and T-J. Wang, *Composite generalized Laguerre-Legendre spectral method with domain decomposition and its application to Fokker-Planck equation an an infinite channel*, Mathematics of Computation, **78** (2008), 129-151.

[53] M. Gustafsson and S. Holmgren, *An Implementation Framework for Solving High-Dimensional PDEs on Massively Parallel Computers*, Numerical Mathematics and Advanced Applications, (2010), 417-424. In G. Kreiss et al. (eds.), Numerical Mathematics and Advanced Applications 2009, Springer-Verlag, Berlin, Heidelberg, 2010.

[54] T. Hillen and H. G. Othmer, *The diffusion limit of transport equations derived from velocity-jump processes*, SIAM J. Appl. Math., **61** (2000), 751-775.

[55] H. Horsthemke, *Spatial instabilities in reaction random walks with direction independent kinetics*, Phys. Rev. E, **60** (1999), 2651-2663.

[56] W. Hundsdorfer, *Numerical Solution of Advection-Diffusion-Reaction Equations.* Lecture Notes, Thomas Stieltjes Institute, 2000.

[57] R. Indira, M. C. Valsakumar, K. P. N. Murthy, and G. Ananthakrishna , *Diffusion in a bistable potential: A comparative study of different methods of solution*, Journal of Statistical Physics, **33** (1983), 181-194.

[58] K. Ishige and M. Murata, *Uniqueness of nonnegative solutions of the Cauchy problem for parabolic equations on manifolds or domains*, Annali della Scuola Normale Superiore di Pisa, Classe di Scienze 4e serie, **30** (2001), 171-223.

[59] K. Ito and K. Kunisch, *Optimal control of parabolic variational inequalities*, J. Math. Pures Appl., **93** (2010), 329-360.

[60] H. G. Jhang and C. S. Chang, *A self-adjoint form of linearized Coulomb collision operator for energetic ions*, Phys. Plasmas, **2** (1995), 3917.

[61] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, Springer, 2010.

[62] N. V. Krylov, *Introduction to the Theory of Random Processes*, American mathematical Society, 2002.

[63] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic type*, American Mathematical Society, vol. 23, Providence, Rhode Island, 1968.

[64] E. W. Larsen, and C. D. Levermore, and G. C. Pomraning, and J. G. Sanderson, *Discretization Methods for One-Dimensional Fokker-Planck Operators.* Journal of Computational Physics, **61** (1985), 359-390.

[65] P. D. Lax, *Hyperbolic Partial Differential Equations-Courant Lecture Notes in Mathematics.* Courant institute of Mathematical Sciences, American Mathematical Society, Providence, RI, (2006).

[66] P. D. Lax and R. D. Richtmyer, *Survey of the stability of linear finite difference equations.* Comm. Pure Appl. Math., **9** (1956), 267-293.

[67] C. Le Bris and P.-L. Lions, *Existence and Uniqueness of Solutions to Fokker-Planck Type Equations with Irregular Coefficients*, Communications in Partial Differential Equations, **33** (2008), 1272-1317.

[68] M. Lefebvre, *Applied Stochastic Processes*, Springer, 2007.

[69] F. Lemeire, *Bounds for condition numbers of triangular and trapezoid matrices*, BIT, **15** (1975) 58-64.

[70] I. L'Heureux, *Reaction rate kernel for dichotomous noise-induced transitions in bistable systems*, Phys. Rev. E, **51** (1995), 2787.

[71] O. Lies-Svendsen and M. H. Rees, *An improved kinetic model for the polar outflow of a minor ion*, Journal of Geophysical Research: Space Physics , **101** (1996), 2415-2433.

[72] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[73] M. van Loon, *Numerical methods in smog prediction.* Master thesis, University of Amsterdam, 1996.

[74] X. Luo and S. S.-T. Yau, *Hermite spectral method to 1D forward Kolmogorov equation and its application to nonlinear filtering problems*, IEEE Trans. Automat. Control, **58** (2013), 2495-2507.

[75] H. Ma, W. Sun and T. Tang, *Hermite spectral methods with a time-dependent scaling for parabolic equations in unbounded domains*, SIAM I. Numer. Anal., **43** (2005), 58-75.

[76] L. Magni, D.M. Raimondo, and F. Allgöwer, *Nonlinear Model Predictive Control*, Springer, Berlin, 2009.

[77] A.N. Malakhov and A.L. Pankratov , *Exact solution of Kramers' problem for piecewise parabolic potential profiles*, Physica A: Statistical Mechanics and its Applications, **229** (1996), 109-126.

[78] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Program, **16** (1979), 98-110.

[79] D.Q. Mayne and H. Michalska, *Receding horizon control for nonlinear systems*, IEEE Trans. Aut. Control, **35** (1990), 814-824.

[80] B. McNamara and Kurt Wiesenfeld, *Theory of stochastic resonance*, Phys. Rev. A, **39** (1989), 4854.

[81] G. N. Mil'shtein and Y. M. Repin, *Action of a Markov process on a system of differential equations*, Differ. Equ. (translated from Russian), **5** (1972), 1010-1019.

[82] A. Morita, *Free Brownian motion of a particle driven by a dichotomous random force*, Phys. Rev. A, **41** (1990), 754-760.

[83] K. W. Morton and D. F. Mayers, *Numerical Solution of Partial Differential Equations: An Introduction*, Cambridge University Press, 2005.

[84] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[85] B. Nowakowski, *Solution of the Fokker-Planck equation for reactive Rayleigh gas*, Phys. Rev. E, **53** (1996), 2964.

[86] Y. Ou and E. Schuster, *On the stability of receding horizon control of bilinear parabolic PDE systems*, Proceedings of the 2010 IEEE Conference on Decision and Control, Atlanta, Georgia, December 15-17, 2010.

[87] H. F. Ouyang, Z. Q. Huang, and E. J. Ding, *1/f noise and one-dimensional Brownian motion in a singular potential*, Phys. Rev. E, **50** (1994), 2491.

[88] B. T. Park, V. Petrosian, *Fokker-Planck equations of stochastic acceleration: A study of numerical methods*, Astrophys. J., Suppl. Ser., **103** (1996), 255-267.

[89] R. F. Pawula and O. Rice, *On filtered binary processes*, IEEE Trans. Inf. Th., **32** (1986), 63-72.

[90] S. Primak, and V. Kontorovich, and V. Lyandres, *Stochastic Methods and Their Applications to Communications.* John Wiley & Sons, 2004.

[91] R. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications.* Springer, Berlin, 1996.

[92] H. Risken and Th. Leiber, *Decay rates for a class of bistable potentials: Parabolic to wedge-shaped form*, Phys. Rev. A, **40** (1989), 1582.

[93] H. Schättler and U. Ledzewicz, *Geometric Optimal Control - Theory, Methods and Examples*, Springer, 2012.

[94] D.F. Shanno, *Conjugate gradient methods with inexact searches*, Math. Oper. Res., 3 (1978), 244-256.

[95] J. Shen and L-L. Wang, *Some recent advances on spectral methods for unbounded domains*, Communications in computational Physics, **5** (2009), 195-241.

[96] J. Shen, *Stable and efficient spectral methods in unbounded domains using Laguerre functions*, SIAM J. NUMER. ANAL., **38** (2000), 1113-1133.

[97] J. Shen, T. Tang and L-L. Wang, *Spectral methods: Algorithms, Analysis and Applications*, Springer, 2011.

[98] B. Shizgal and R. Blackmore, *Discrete ordinate method of solution of a Fokker-Planck equation with a bistable potential*, Chemical Physics Letters, **109** (1984), 242-245.

[99] B. Shizgal and D. R. A. McMahon, *Electric field dependence of transient electron transport properties in rare-gas moderators*, Phys. Rev. A, **32** (1985), 3669.

[100] V. A. Shneidman, *Comment on "Transient kinetics of nucleation"*, Phys. Rev. A, **44** (1991), 8441.

[101] S. Stepanow, *Kramers equation as a model for semiflexible polymers*, Phys. Rev. E, **54** (1996), R2209(R).

[102] E. Süli, *Numerical Solution of PDEs.* Lecture Notes, Oxford University, 2005.

[103] T. Theuns, *Numerical study of energy diffusion in King models*, Mon. Not. R. Astron. Soc., **279** (1996), 827-836.

[104] F. Tröltzsch, *Optimal Control of Partial differential Equations: Theory, Method, and Applications*, AMS, 2010.

[105] R.S. Varga, *Matrix Iterative Analysis.* Springer, 2000.

[106] P. Wilmott, S. Howison, and J. Dewynne, *The Mathematics of Financial Derivatives*, Cambridge University Press, 1995.

[107] X.-G. Wu and R. Kapral, *Projected Dynamics: Analysis of a Chemical Reaction Model*, J. Chem. Phys., **91** (1989), 5528-5543.

[108] K. R. Yawn, B. N. Miller, and W. Maier, *Stochastic dynamics of gravity in one dimension*, Phys. Rev. E, **52** (1995), 3390.

[109] A. Zisowsky and M. Ehrhardt, *Discrete transparent boundary conditions for parabolic systems*, Mathematical and Computer Modelling, **43** (2006), 294-309.

[110] H. bin Zubair, C.C. Oosterlee, and R. Wienands, *Multigrid for high dimensional elliptic partial differential equations on non-equidistant grid*, SIAM J. Sci. Comput., **29** (2007), 1613-1636.