Thesis submitted for the degree of
Doctor of Natural Sciences

Theoretical and numerical analysis of Fokker–Planck optimal control problems by first– and second–order optimality conditions

Jacob Körner

Würzburg, June 3, 2024

Julius-Maximilians-University of Würzburg
Institute for Mathematics, Chair of Scientific Computing
Advisor: Prof. Dr. Alfio Borzì

Acknowledgements

I would like to express my deepest gratitude to my supervisor Alfio Borzì for his expertise, encouraging words, and for navigating me through the difficulties of this thesis. Due to his wise guidance, I was able to keep my eye on the actual goal over the three years of research.

Many thanks to all of my colleagues at the chair of Scientific Computing for the positive and friendly working atmosphere. Special thanks to Daniel Wachsmuth, Frank Werner and Markus Weimar, who always managed to find solutions or satisfactory answers to all my questions on differential equations and optimization problems throughout the last three years. I am very grateful to have had Jan Bartsch and Nadja Vater as my PhD colleagues, as they helped improve my numerical codes, my research work and my lunch breaks significantly.

I would like to acknowledge my former Professor Gerhard Rein and classmates from my time at the University of Bayreuth. Special thanks to my friend Nadine Ansorge for encouraging me to start a doctorate, for always cheering me up when I needed it and being on my side for so long.

Words cannot express my gratitude to my parents, Eva and Hermann, to my brother Max and to my grandparents Waltraud and Fritz for the lifelong and unconditional support.

Würzburg, January 2024 *Jacob Körner*

Abstract

In this thesis, a variety of Fokker–Planck (FP) optimal control problems are investigated. Main emphasis is put on a first– and second–order analysis of different optimal control problems, characterizing optimal controls, establishing regularity results for optimal controls, and providing a numerical analysis for a Galerkin–based numerical scheme.

The Fokker–Planck equation is a partial differential equation (PDE) of linear parabolic type deeply connected to the theory of stochastic processes and stochastic differential equations. In essence, it describes the evolution over time of the probability distribution of the state of an object or system of objects under the influence of both deterministic and stochastic forces. The FP equation is a cornerstone in understanding and modeling phenomena ranging from the diffusion and motion of molecules in a fluid to the fluctuations in financial markets.

Two different types of optimal control problems are analyzed in this thesis. On the one hand, Fokker–Planck ensemble optimal control problems are considered that have a wide range of applications in controlling a system of multiple non–interacting objects. In this framework, the goal is to collectively drive each object into a desired state. On the other hand, tracking–type control problems are investigated, commonly used in parameter identification problems or stemming from the field of inverse problems. In this framework, the aim is to determine certain parameters or functions of the FP equation, such that the resulting probability distribution function takes a desired form, possibly observed by measurements. In both cases, we consider FP models where the control functions are part of the drift, arising only from the deterministic forces of the system. Therefore, the FP optimal control problem has a bilinear control structure. Box constraints on the controls may be present, and the focus is on time–space dependent controls for ensemble–type problems and on only time–dependent controls for tracking–type optimal control problems.

In the first chapter of the thesis, a proof of the connection between the FP equation and stochastic differential equations is provided. Additionally, stochastic optimal control problems, aiming to minimize an expected cost value, are introduced, and the corresponding formulation within a deterministic FP control framework is established. For the analysis of this PDE–constrained optimal control problem, the existence, and regularity of solutions to the FP problem are investigated. New L^∞–estimates for solutions are established for low space dimensions under mild assumptions on the drift. Furthermore, based on the theory of Bessel potential spaces, new smoothness properties are derived for solutions to the FP problem in the case of only time–dependent controls. Due to these properties, the control–to–state map, which associates the control functions with the corresponding solution of the FP problem, is well–defined, Fréchet differentiable and compact for suitable Lebesgue spaces or Sobolev spaces. The existence of optimal controls is proven under various assumptions on the space of admissible controls and objective functionals. First–order optimality conditions are derived using the adjoint system. The resulting characterization of optimal controls is exploited to achieve higher regularity of optimal controls, as well as their state and co–state functions. Since the FP optimal control problem is non–convex due to its bilinear structure, a first–order analysis should be complemented by a second–order analysis. Therefore,

a second–order analysis for the ensemble–type control problem in the case of H^1–controls in time and space is performed, and sufficient second–order conditions are provided. Analogous results are obtained for the tracking–type problem for only time–dependent controls.

The developed theory on the control problem and the first– and second–order optimality conditions is applied to perform a numerical analysis for a Galerkin discretization of the FP optimal control problem. The main focus is on tracking-type problems with only time–dependent controls. The idea of the presented Galerkin scheme is to first approximate the PDE–constrained optimization problem by a system of ODE–constrained optimization problems. Then, conditions on the problem are presented such that the convergence of optimal controls from one problem to the other can be guaranteed. For this purpose, a class of bilinear ODE–constrained optimal control problems arising from the Galerkin discretization of the FP problem is analyzed. First– and second–order optimality conditions are established, and a numerical analysis is performed. A discretization with linear finite elements for the state and co–state problem is investigated, while the control functions are approximated by piecewise constant or piecewise quadratic continuous polynomials. The latter choice is motivated by the bilinear structure of the optimal control problem, allowing to overcome the discrepancies between a discretize–then–optimize and optimize–then–discretize approach. Moreover, second–order accuracy results are shown using the space of continuous, piecewise quadratic polynomials as the discrete space of controls. Lastly, the theoretical results and the second–order convergence rates are numerically verified.

Zusammenfassung

In dieser Dissertation werden verschiedene Fokker–Planck (FP) optimale Steuerungsprobleme untersucht. Die Schwerpunkte liegen auf einer Analyse von Optimalitätsbedingungen erster und zweiter Ordnung, der Charakterisierung optimaler Steuerungen, dem Herleiten höhere Regularität von optimalen Kontrollen sowie einer theoretischen numerischen Analyse für ein numerisches Verfahren basierend auf einer Galerkin Approximation.

Die Fokker–Planck Gleichung ist eine lineare, parabolische, partielle Differentialgleichung (PDE), die aus dem Gebiet stochastischer Differentialgleichungen und stochastischer Prozesse stammt. Im Wesentlichen beschreibt sie die zeitliche Entwicklung der Wahrscheinlichkeitsverteilung des Zustands eines Objekts bzw. eines Systems von Objekten unter dem Einfluss sowohl deterministischer als auch stochastischer Kräfte. Die Fokker–Planck Gleichung ist ein Eckpfeiler zum Verständnis und Modellieren von Phänomenen, die von der Diffusion und Bewegung von Molekülen in einer Flüssigkeit bis hin zu den Schwankungen in Finanzmärkten reichen.

Zwei verschiedene Arten von optimalen Kontrollproblemen werden in dieser Arbeit umfassend analysiert. Einerseits werden Fokker–Planck Ensemble Steuerungsprobleme betrachtet, die in der Kontrolle von Systemen mit mehreren nicht wechselwirkenden Objekten vielfältige Anwendungen haben. In diesem Gebiet ist das Ziel, alle Objekte gemeinsam in einen gewünschten Zustand zu lenken. Andererseits werden Tracking Kontrollprobleme untersucht, die häufig bei Parameteridentifikationsproblemen auftreten oder aus dem Bereich inverser Probleme stammen. Hier besteht das Ziel darin, bestimmte Parameter oder Funktionen der Fokker–Planck Gleichung derart zu bestimmen, dass die resultierende Wahrscheinlichkeitsverteilung eine gewünschte Form annimmt, welche beispielsweise durch Messungen beobachtet wurde. In beiden Fällen betrachten wir FP Modelle, bei denen die Kontrollfunktion Teil des sogenannten Drifts ist, das heißt der Teil, der nur aus den deterministischen Kräften des Systems resultiert. Daher hat das FP Kontrollproblem eine bilineare Struktur. Untere und obere Schranken für die Kontrollfunktionen können vorhanden sein, und der Fokus liegt auf zeit– und raumabhängigen Steuerungen für Ensemble Kontrollprobleme, sowie auf nur zeitlich abhängigen Steuerungen für Tracking Kontrollprobleme.

Am Anfang der Dissertation wird ein Beweis für den Zusammenhang zwischen der FP Gleichung und stochastischen Differentialgleichungen dargelegt. Darüber hinaus werden stochastische optimale Steuerungsprobleme eingeführt, deren Ziel es ist, einen erwarteten Kostenwert zu minimieren. Zusätzlich wird das Problem als ein deterministisches FP Kontrollproblem formuliert. Für die Analyse dieses Kontrollproblems wird die Existenz und Regularität von Lösungen für die FP Differentialgleichung untersucht. Neue L^∞–Abschätzungen für Lösungen werden für niedrige Raumdimensionen unter schwachen Annahmen an den Drift bewiesen. Zusätzlich werden, basierend auf der Theorie über Bessel Potentialräume, neue Glattheitseigenschaften für Lösungen des FP–Problems im Falle zeitabhängiger Steuerungen erarbeitet. Aufgrund dieser Eigenschaften ist die sogenannte control–to–state Abbildung, welche die Kontrollfunktion mit der entsprechenden Lösung des FP Problems verknüpft, wohldefiniert, Fréchet–differenzierbar und kompakt für geeignete Lebesgue–Räume oder Sobolev–Räume.

Die Existenz optimaler Steuerungen wird unter verschiedenen Annahmen an den Funktionenraum der

Kontrollen und des Kostenfunktionals bewiesen. Optimalitätsbedingungen erster Ordnung werden unter Verwendung des adjungierten Systems aufgestellt. Die daraus resultierende Charakterisierung optimaler Steuerungen wird genutzt, um eine höhere Regularität optimaler Steuerungen sowie ihrer Zustandsfunktion und des adjungierten Problems zu erhalten. Da das FP Kontrollproblem aufgrund der bilinearen Struktur nicht konvex ist, sollte eine Analyse von Optimalitätsbedingungen erster Ordnung durch eine Analyse von Optimalitätsbedingungen zweiter Ordnung ergänzt werden. Dies wird für das Ensemble Kontrollproblem im Fall von zeit– und ortsabhängigen Steuerungen mit H^1–Regularität durchgeführt, und hinreichende Bedingungen für lokale Minimierer werden hergeleitet. Analoge Ergebnisse werden für das Tracking–Problem für nur zeitabhängige Steuerungen bewiesen.

Die entwickelte Theorie zu diesem optimalen Steuerungsproblem und dessen Optimalitätsbedingungen wird angewendet, um eine numerische Analyse für eine Galerkin–Diskretisierung des FP Kontrollproblems durchzuführen. Der Schwerpunkt liegt auf Tracking–Problemen mit nur zeitabhängigen Steuerungen. Die Idee des vorgestellten Galerkin–Verfahrens besteht darin, das PDE–Optimierungsproblem zunächst durch ein System von Optimierungsproblemen mit gewöhnlichen Differentialgleichungen (ODE) als Nebenbedingung zu approximieren. Dann werden Bedingungen an das Problem präsentiert, sodass die Konvergenz optimaler Steuerungen von einem Problem zum anderen garantiert werden kann. Zu diesem Zweck wird eine Klasse bilinearer ODE–Kontrollprobleme analysiert, welche sich aus der Galerkin–Diskretisierung des FP Problems ergeben. Optimalitätsbedingungen erster und zweiter Ordnung werden bewiesen, und eine numerische Analyse wird durchgeführt. Eine Diskretisierung mit linearen Finiten–Elementen der Zustands– und Adjungiertengleichung wird untersucht, während die Kontrollfunktionen durch stückweise konstante oder stetige, stückweise quadratische Polynome approximiert werden. Diese Wahl wird durch die bilineare Struktur des optimalen Kontrollproblems begründet, da sie es ermöglicht, die Diskrepanzen zwischen einem Ansatz von „zuerst diskretisieren dann optimieren” und „zuerst optimieren dann diskretisieren” zu überwinden. Durch die Verwendung stetiger, stückweise quadratischer Polynome als Diskretisierung der Steuerungen kann außerdem quadratische Konvergenzordnung gezeigt werden. Abschließend werden die theoretischen Ergebnisse und die Konvergenzraten zweiter Ordnung numerisch verifiziert.

Contents

1
Introduction

The Fokker-Planck (FP) equation is a fundamental partial differential equation (PDE) that plays a central role in the field of statistical physics, particularly in describing the dynamics of stochastic processes. It finds widespread application in understanding the behavior of (partially) random systems and the evolution of probability distributions associated with them.

Named after physicists Adriaan Fokker and Max Planck, this equation provides a mathematical framework for modeling the probability density function of the motion of a particle or other relevant variables in the presence of random or stochastic forces. Initially derived to analyze the Brownian motion problem, the Fokker–Planck equation addresses scenarios where a small yet macroscopic particle is immersed in a fluid. In this context, the fluid's molecules exert unpredictable kicks on the particle, leading to fluctuations in its velocity. The consequence is an inherent uncertainty regarding the particle's exact dynamics, giving rise to a probability distribution $p = p(t, x)$, where the integral $\int_U p(t, x)\ dx$ gives the probability of the particle having a velocity $\bar{x} \in U$ at time t. As the scientific landscape evolves, nowadays, the application of the Fokker–Planck equation extends beyond its original association with Brownian motion. It has seamlessly integrated into diverse fields within the natural sciences – from solid–state physics to quantum optics, chemical physics, theoretical biology, and circuit theory – the Fokker–Planck equation stands as a versatile tool to analyze complex systems under the influence of random fluctuations. The investigation of the motion of particles has evolved into modelling collective motion of groups, such as the movement of molecules and bacteria, and the motion of herds of animals like fishes and birds. Furthermore, the applications of the Fokker–Planck equation are not limited to the case of probability distributions, or where the state x in the probability distribution function (PDF) $p = p(t, x)$ is the velocity or position of the object. As an example, we mention the Black–Scholes equation as a special case of a FP equation. This equation appears in mathematical finance to model the price evolution $V = V(t, S)$ of derivatives under the Black–Scholes model, and in that context, V is a function of time t and the stock price S and it is not a PDF. Another relevant example stems from the field of stochastic epidemic models, where it is the aim to model the spread of diseases among a population of size N, which is divided into k different compartments $x = (x_1, \ldots, x_k)$. One individual belongs at one point in time t exactly to one compartment x_i, and the individual will eventually transfers to other compartments due to the dynamic of the disease. Typical example of compartments in epidemic models are, among others, the group of healthy and susceptible individuals, the group of infectious individuals and the group of recovered individuals. In that context, integrating the corresponding probability distribution function p, given by the FP equation, over a region $U \subset [0, N]^k$ at time t, yields the probability of $(x_1, \dots, x_k)$ taking values in U.

A general concept in modelling complex systems in natural sciences is that once a suitable mathematical framework is established, we are interested in controlling the possible outcomes. In terms of the FP equation, this means the following. The FP equation answers the question, how the possibility of certain outcomes evolve over time, under the assumption that the dynamics of the problem are known. The question of controlling a process goes the other way around. Thus, we ask, how we have to adapt the dynamics of the problem, such that certain desired outcomes are highly likely; or certain undesired outcomes are very unlikely, respectively. If we go back to the Brownian motion problem formulated in a Fokker–Planck control framework, we are no longer interested in the motion that the particle will most likely have given its force field, but we want to determine a specific force field such that the particle has (most likely) the desired motion. In the context of stochastic epidemiological models and the modelling of infectious diseases, the typical controllable dynamics of the system are, among others, vaccination of the individuals or reducing the contact rate between individuals. Therefore, a Fokker–Planck control framework provides a robust mathematical tool to analyze which of these actions to take, in order to reduce the likelihood of an outbreak of the disease.

In summary, the Fokker–Planck equation and the Fokker–Planck optimal control framework are powerful tools providing a formalism to study and control the dynamics of systems subject to random influences. Their application extends across various scientific disciplines, making it a key concept in the analysis and the control of stochastic processes and their impact on the behavior of physical, financial and biological systems.

The thesis is organized as follows. In Chapter 1, we introduce stochastic differential equations and recall basic definitions such as the Itô integral and the Wiener processes. Then, we prove that the evolution of probability distribution functions for certain stochastic processes is given by the solution of the Fokker–Planck problem. Based on this, the connection of stochastic optimal control problems and Fokker–Planck optimal control problems is established. In order to analyze the resulting PDE–constrained optimal control problem, we consider optimization problems and useful optimality conditions in a general setting, while main emphasis is put on problems formulated in infinite–dimensional Banach spaces.

The second chapter is devoted to the analysis of the (inhomogeneous) Fokker–Planck differential equation with flux–zero boundary conditions and given initial data. We introduce the concept of weak solutions, prove uniqueness and well–posedness of the problem, and derive higher regularity of solutions under additional assumptions. Furthermore, a result on maximal L^p–regularity is given, based on the theory of Besov spaces. Next, the control–to–state map is introduced and its well–posedness on a variety of sets of admissible controls is discussed. We proceed by proving Fréchet differentiability, Lipschitz continuity and compactness results. The second chapter is concluded by analyzing the linearized Fokker–Planck problem.

In Chapter 3, a variety of ensemble FP optimal control problems are investigated. We prove existence of optimal controls in different settings and also discuss under which conditions the ensemble optimal control problem does not possess solutions. Afterward, we perform a first–order analysis by an adjoint–based approach and provide implicit representations for optimal controls. The chapter is closed with a second–order analysis, relying on the general theory of optimization problems in Banach spaces developed in Section 1.3.

Chapter 4 is devoted to the FP tracking optimal control problem. We focus on only time–dependent, vector valued controls, motivated by the ansatz $u(t,x) = M(x)\tilde{u}(t)$. This ansatz is a reasonable trade–off between complexity and accuracy of the problem, assuming that the space dependency of u can be represented sufficiently well by the components of the (possibly high–dimensional) matrix–valued function $M = M(x)$. Similarly to the previous chapter, we perform a First– and second–order analysis. The results are essential for the detailed numerical analysis performed in Chapters 5–7.

In Chapter 5, we formulate our discretization scheme in a general setting. The idea is to first approximate the PDE optimal control problem by a semidiscrete Galerkin scheme, which yields a sequence of ODE–constrained optimal control problems. Subsequently, we provide conditions on the problem such that convergence of optimal controls from the ODE–optimal control problem to the PDE optimal control holds. Then, we apply this method to our FP control problem and derive the corresponding ODE–control problem. In Chapter 6, this problem is analyzed in depth, and a numerical analysis for a finite element discretization is provided. Linear and quadratic convergence rates are proven, and a numerical test is performed to validate the results.

In Chapter 7 of this thesis, the findings of Chapters 4–6 are combined, and convergence rates of numerical solutions to the Fokker–Planck optimal control problem of tracking type are established.

The results of this thesis are based, in part, on research for the following scientific papers:

J. Körner and A. Borzì, *Second–order analysis of Fokker–Planck ensemble optimal control problems*, ESAIM: Control, Optimisation and Calculus of Variations, (2022).

J. Körner and A. Borzì, *Accuracy estimates for bilinear optimal control problems governed by ordinary differential equations*, Numerical Functional Analysis and Optimization, 44 (2023), p. 564–602.

J. Körner and A. Borzì, *Accuracy of semidiscrete Galerkin approximations to optimal control problems with an application to the Fokker–Planck problem*, submitted to Journal of Dynamical and Control Systems.

Notations

We use the following notations throughout this thesis. Let $\mathbb{N} = \{1, 2, \ldots\}$ denote the set of natural numbers without zero and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}$ and two vectors $x, y \in \mathbb{R}^n$, we denote by

$$x \cdot y := x^\top y = \sum_{i=1}^{n} x_i y_i, \quad |x| := \sqrt{x \cdot x}$$

the Euclidean scalar product and the norm, where $x^\top$ is its transposed. An inequality between two vectors x, y or between a vector and a number $z \in \mathbb{R}$, is to be understood componentwisely, i.e., $x \le y$ iff $x_i \le y_i$ for all $i = 1, \ldots, n$, and $x \le z$ iff $x_i \le z$ for all $i = 1, \ldots, n$. When we write $a > 0$, this shall always imply that a is a real number. Given any set $M \subset \mathbb{R}^n$, we denote by $C(M) = C^0(M)$ the set of continuous functions from M to $\mathbb{R}$. For functions $f : M \to \mathbb{R}$, we define the sup norm as follows

$$\|f\|_\infty = \sup\{|f(x)| : x \in M\}.$$

Let U be an open set and $k \in \mathbb{N}_0 \cup \{\infty\}$. We denote by $C^k(U)$ the space of all k–times continuously differentiable functions on U with norm

$$\|f\|_{C^k(U)} := \sum_{|\alpha|_1 \le k} \|D^\alpha f\|_\infty, \quad k \ne \infty.$$

In that context, $\alpha \in \mathbb{N}_0^n$ is a multi–index with $|\alpha|_1 := \sum_{i=1}^n \alpha_i$ and the corresponding derivative reads

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}}.$$

For $s \in]0,1[$, we denote by $C^s(U) = C^{0,s}(U)$ the space of all Hölder continuous functions to the exponent s with the semi–norm

$$\|f\|_{C^s(U)} := \sup\left\{ \frac{|f(x)-f(y)|}{|x-y|^s} : x, y \in U \text{ and } x \neq y \right\}.$$

The Banach space $C^{k+s}(U) = C^{k,s}(U)$ for $0 < s < 1$ is the set of all functions which have Hölder continuous k–th derivatives, and we introduce the corresponding norm

$$\|f\|_{C^{k+s}(U)} := \sum_{|\alpha|_1 \leq k} \left(\|D^\alpha f\|_\infty + \|D^\alpha f\|_{C^s(U)} \right).$$

We mention that we may abbreviate the following words throughout the thesis: subject to (s.t.), almost everywhere (a.e.), for almost every (f.a.e.) and with respect to (w.r.t.).
For any measurable $f : U \to \mathbb{R}$, we denote by $f_+ := \max\{f, 0\}$ and $f_- := \min\{f, 0\}$ its positive and negative part, respectively. Furthermore, we define the support of f as follows

$$\operatorname{supp} f := U \setminus \bigcup \big\{ U' \subset U \text{ open } : f_{|U'} = 0 \text{ a.e.} \big\}.$$

The symbol $C_c^k(U)$ denotes the space of functions from $C^k(U)$ with compact support in U.
Next, we introduce the Lebesgue spaces for $p \in [1, \infty]$

$$L^p(U) := \{ f : U \to \mathbb{R} \text{ measurable } : \|f\|_{L^p(U)} < \infty \}.$$

As usual, two functions belonging to $L^p(U)$ which agree almost everywhere are identified, and thus, we may introduce the (full) norms

$$\|f\|_{L^q(U)} := \left(\int_U |f(x)|^q \, dx \right)^{1/q}, \quad q \in [1, \infty[,$$
$$\|f\|_{L^\infty(U)} := \inf \big\{ L \geq 0 : |f(x)| \leq L \quad \text{f.a.e. } x \in U \big\}.$$

If clear from the context, we will write $\|f\|_p := \|f\|_{L^p(U)}$ for $p \in [1, \infty]$; notice that the L^∞–norm and sup–norm coincide for pointwisely defined functions on U. We remark that pointwisely defined operations for functions from Lebesgue spaces are, in general, not well–defined. Thus, throughout this thesis, we will always mention it clearly if a pointwisely defined representative of some function $f \in L^p(U)$ is chosen. If such function f is known to be continuous (after possibly modifying f on a set of measure zero), we will write $f \in L^p(U) \cap C(U)$ in order to choose the continuous representative. For $k \in \mathbb{N}_0$ and $p \in [1, \infty]$, we introduce the Sobolev spaces $W^{k,p}(U)$ consisting of all functions $f : U \to \mathbb{R}$ with weak derivatives $D^\alpha f \in L^p(U)$, $|\alpha|_1 \leq k$. The corresponding norms are given by

$$\|f\|_{W^{k,q}(U)} := \left(\sum_{|\alpha|_1 \leq k} \|D^\alpha f\|_q^q \right)^{1/q}, \quad 1 \leq q < \infty,$$
$$\|f\|_{W^{k,\infty}(U)} := \max_{|\alpha|_1 \leq k} \|D^\alpha f\|_\infty.$$

For an open, non–empty interval $]a, b[$, we write $L^p(a, b) := L^p(\,]a, b[\,)$ with analogous notations for Sobolev spaces. Given any Banach space $(X, \|\cdot\|_X)$ and a subspace $Y \subset X$, we define

$$\overline{Y}^X := \big\{ g \in X : \text{ there exists } (g_j)_{j \in \mathbb{N}} \subset Y \text{ such that } \|g - g_j\|_X \to 0 \text{ as } j \to \infty \big\}$$

as the closure of Y w.r.t. X. In that context, we recall that $W_0^{k,p} := \overline{C_c^\infty(U)}^{W^{k,p}(U)}$. For the special case $p = 2$, we use the common notation for the Hilbert space $H^k(U) := W^{k,2}(U)$ and $H_0^k(U) := W_0^{k,2}(U)$

with scalar product

$$\langle f,g\rangle_2 := \langle f,g\rangle_{L^2(U)} := \int_U f(x)g(x)\,dx, \quad \langle f,g\rangle_{H^k} := \sum_{|\alpha|_1\leq k} \langle D^\alpha f, D^\alpha g\rangle_2.$$

Throughout this thesis, the variable $T>0$ denotes a final time horizon. Let $(X,\|\cdot\|_X)$ denote a separable Banach space. Then, we introduce the Bochner space and corresponding norm

$$C([0,T];X) := \{u:[0,T]\to X : u \text{ is continuous }\}, \quad \|u\|_{C([0,T];X)} = \max_{t\in[0,T]} \|u(t)\|_X, \tag{1.1}$$

with analogous definitions for $C^k([0,T];X)$, $k\in\mathbb{N}$. For $p\in[1,\infty]$, we define the following Banach space

$$L^p(0,T;X) := \{u:[0,T]\to X : u \text{ measurable and } \|u(\cdot)\|_X \in L^p(0,T)\}$$

with corresponding norms

$$\begin{aligned}\|u\|_{L^q(0,T;X)} &:= \left(\int_0^T \|u(t)\|_X^q\,dt\right)^{1/q}, \quad q\in[1,\infty[,\\ \|u\|_{L^\infty(0,T;X)} &:= \operatorname*{ess\,sup}_{t\in[0,T]} \|u(t)\|_X.\end{aligned}$$

Sobolev spaces with values in Banach spaces are defined as

$$W^{1,p}(0,T;X) := \{u\in L^p(0,T;X) : \partial_t u \in L^p(0,T;X)\}, \quad p\in[1,\infty]$$

with norms

$$\begin{aligned}\|u\|_{W^{1,q}(0,T;X)} &:= \left(\|u\|^q_{L^q(0,T;X)} + \|\partial_t u\|^q_{L^q(0,T;X)}\right)^{1/q}, \quad q\in[1,\infty[,\\ \|u\|_{W^{1,\infty}(0,T;X)} &:= \max\left\{\|u\|_{L^\infty(0,T;X)}, \|\partial_t u\|_{L^\infty(0,T;X)}\right\}.\end{aligned}$$

For any space X and $n\in\mathbb{N}$, we say that $f\in X^n$ if every component f_i of f belongs to X for $i=1,\dots,n$. In that sense, we remark that $L^p(U;\mathbb{R}^n) = L^p(U)^n$ with analogous definitions for Sobolev spaces. Additionally, we define the L^p–norm of a vector valued function $f\in L^p(\Omega)^m$ for later conveniences as

$$\|f\|_{L^q(\Omega)^m} = \|f\|_q = \left(\sum_{i=1}^m \|f_i\|_q^q\right)^{1/q}, \quad 1\leq q<\infty, \qquad \|f\|_{L^\infty(\Omega)^m} = \|f\|_\infty = \sum_{i=1}^m \|f_i\|_\infty.$$

Notice that the norms $\|f\|_p$ and $\||f|\|_p$ are equivalent on $L^p(U)$ but have different values in general. Throughout this thesis, when we consider the Fokker–Planck problem, we use $d\in\mathbb{N}$ as the spatial dimension, and $\Omega\subset\mathbb{R}^d$ denotes a convex domain that is polygonal or has sufficiently smooth boundary $\partial\Omega$. The time–space cylinder is denoted by $\Omega_T :=]0,T[\times\Omega$. For functions defined on Ω_T, we write ∂_t for the classical, weak or distributional time derivative, and ∂_{x_i} denotes the classical, weak or distributional derivative w.r.t. x_i for $i=1,\dots,d$. Furthermore, we write for $u:\Omega_T\to\mathbb{R}$

$$\dot u = \partial_t u, \quad \nabla u = \nabla_x u = (\partial_{x_1},\dots,\partial_{x_d})u, \quad Du = (\partial_t,\partial_{x_1},\dots,\partial_{x_d})u.$$

For m–dimensional vector valued functions $u:\Omega_T\to\mathbb{R}^d$, we may interpret for convenience ∇u or $\mathrm{D}u$ as md or $m(d+1)$ dimensional vector, respectively. The divergence of a vector field $u=(u_1,\dots,u_d)$ is denoted by $\operatorname{div} u := \nabla\cdot u = \sum_{i=1}^d \partial_{x_i} u_i$. Lastly, integrals and the dependencies of functions can be abbreviated if the dependencies are clear from the context. As an example, for f defined on Ω_T, we may write $\int_{\Omega_T} f(t,x)\,dt\,dx = \int_{\Omega_T} f\,dt\,dx$ or $\int_\Omega f(t,x)\,dx = \int_\Omega f(t)\,dx$. In addition, we use the common notation with dots to emphasize the dependence of functions or operators, for example the notation $f(t) = f(t,\cdot)$ interprets $f:\Omega_T\to\mathbb{R}$ as a function defined on Ω for some fixed $t\in]0,T[$.

1.1 Particles under uncertainty – a derivation of the Fokker–Planck equation arising from stochastic differential equations

> *These motions were such as to satisfy me, after frequently repeated observation, that they arose neither from currents in the fluid, nor from its gradual evaporation, but belonged to the particle itself.*
>
> Robert Brown, 1773 – 1859

Consider a large particle with mass M suspended in a liquid or a gas medium consisting of smaller particles with lighter mass $m \ll M$. We assume the only force on the large particle is given by an exterior force field. Our aim is to compute the motion of the large particle through the medium, taking into account potential collisions with smaller particles that may affect its path and velocity.

In a deterministic setting, where the velocity and position of each small particle are known precisely at each time, it is possible – at least theoretically in the framework of classical mechanics – to keep track of each collision and calculate the trajectory of the large particle. However, when we consider the scale of molecules, up to 10^{20} collisions can occur during one second, depending on the temperature of the medium, and therefore, a statistical approach is necessary.

Many scientists have investigated this topic for a long time. The first one who discovered and described this random motion of particles was the botanist Robert Brown in 1827 [17]. The first rigorous formulation of it via stochastic processes and the concept of Brownian motion was derived in 1900 by Louis Jean–Baptiste Alphonse Bachelier. About five years later, Albert Einstein [31] and Marian Smoluchowski [65] studied this problem in the framework of statistical mechanics, which attracted a lot of interest from the physics community. The list of famous physicists and mathematicians that were involved in the many scientific breakthroughs is long and for a compelling overview on Brownian motion and its statistical description, we refer to [32, 54].

It is our aim to introduce the mathematical formalism of a motion of such Brownian particles in the context of stochastic differential equations (SDEs). Furthermore, we investigate its path with a probability distribution function (PDF) which yields a rigorous derivation of the FP equation. For this purpose, we introduce some basic definitions and properties of probability theory, SDEs and stochastic processes; for more details and the proofs we refer to the books [37, 52].

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space. With $\mathcal{B}^d$ we denote the Borel σ–algebra, that is, the σ–algebra generated by the family of all open sets in $\mathbb{R}^d$. We say that two random variables $X, Y : \Omega \to \mathbb{R}^d$ are independent if for all $A, B \in \mathcal{B}^d$

$$P\{\omega \in \Omega \mid X(\omega) \in A \text{ and } Y(\omega) \in B\} = P\{\omega \in \Omega \mid X(\omega) \in A\} P\{\omega \in \Omega \mid Y(\omega) \in B\}.$$

We recall that any given random variable $X : \Omega \to \mathbb{R}^d$ induces another probability measure P_X, the distribution of X under P, given by

$$P_X : \mathcal{B}^d \to [0,1], \quad P_X(A) := P\{\omega \in \Omega \mid X(\omega) \in A\}.$$

Further, any measurable function $f : \mathbb{R}^d \to [0, \infty]$ with normed (Lebesgue) integral $\int_{\mathbb{R}^d} f(x)\, dx = 1$ induces a probability measure, given by

$$P_f : \mathcal{B}^d \to [0,1], \quad P_f(A) := \int_A f(x)\, dx, \quad A \in \mathcal{B}^d.$$

In this context, we say that a random variable $X : \Omega \to \mathbb{R}^d$ induces a (Lebesgue) density function or PDF $f : \mathbb{R}^d \to [0, \infty]$ if $P_X = P_f$. For any integrable random variable $X : \Omega \to \mathbb{R}^d$, we denote by $\mathbb{E}[X]$ its expected value. The next lemma is an important tool to calculate the expected value of a (transformed) random variable given its PDF.

Lemma 1.1.1. *Let $X : \Omega \to \mathbb{R}^d$ be a random variable that has a PDF f and let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be measurable. Then, it holds that*

$$\varphi(X) \text{ is integrable} \quad \Longleftrightarrow \quad \int_{\mathbb{R}^d} |\varphi(x)| f(x)\, dx < \infty.$$

In this case we obtain

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}^d} \varphi(x) f(x)\, dx.$$

Notice that the expected value and the variance $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ of a random variable X can be calculated by its PDF using Lemma 1.1.1:

$$\mathbb{E}\left[X^k\right] = \int_{\mathbb{R}^d} x^k f(x)\, dx, \quad k \in \mathbb{N}.$$

In the following, let $\{\mathcal{F}_t\}_{t\geq 0}$ be a filtration of $\mathcal{F}$ that satisfies the usual conditions, that is, $\{\mathcal{F}_t\}_{t\geq 0}$ is a family of increasing sub–σ-algebras of $\mathcal{F}$ and $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$ for all $t \geq 0$.
On some interval $I \subset \mathbb{R}$, a stochastic process is a family $\{X_\cdot(t)\}_{t\in I}$ of $\mathbb{R}^d$–valued random variables. Usually, we consider $I = \mathbb{R}, [0, \infty[$ or $[0, T]$ for some $T > 0$. For fixed outcome $\omega \in \Omega$, we introduce the sample path of the process

$$t \mapsto X_\omega(t) \in \mathbb{R}^d.$$

Similarly, for fixed time $t \in I$, we recall that $\omega \mapsto X_\omega(t) \in \mathbb{R}^d$ is a random variable. Notice that in the context of stochastic processes, we write the argument ω of the random variable $\omega \mapsto X_\omega(t)$ as a lower index. A stochastic process is sometimes also considered as a function of two variables from $I \times \Omega$ to $\mathbb{R}^d$. We say that a stochastic process is continuous if for almost all $\omega \in \Omega$, $t \mapsto X_\omega(t)$ is continuous. It is said to be adapted if for every $t \in I$ the random variable $X(t)$ is $\mathcal{F}_t$ measurable. A stochastic process is integrable if for every $t \in I$, $X(t)$ is an integrable random variable and hence $X(t) \in L^1(\Omega)$.
Now, we can rigorously place the famous Brownian motion within the context of stochastic processes, leading to the definition of a Wiener process.

Definition 1.1.2. *Let $W = \{W_\cdot(t)\}_{t\in I}$ be a real–valued, continuous adapted stochastic process. We say that W is a (standard one–dimensional) Wiener process or a Brownian motion if the following holds:*

(i) $W(0) = 0$ *almost surely;*

(ii) for all $0 \leq s, t < \infty$, *the increment* $W(t + s) - W(t)$ *is normally distributed and*

$$\mathbb{E}[W(t + s) - W(t)] = 0, \quad \mathrm{Var}[W(t + s) - W(t)] = s;$$

(iii) for all $0 \leq s < t < \infty$, $0 < \tau < t$, *the increment* $W(t + s) - W(t)$ *is independent of* $W(\tau)$.

We say $W = (W^1, \ldots, W^d)$ is a d–dimensional Wiener Process if each W^i is a one–dimensional Wiener process and $W^1, \ldots, W^d$ are independent.

With the notion of a Wiener process, we can introduce the Itô integral of a stochastic process X

$$\int_a^b X(s)\, dW(s),$$

defined on the space of all real–valued continuous adapted processes X such that

$$\mathbb{E}\left[\int_a^b |X(t)|^2\, dt\right] < \infty.$$

We refer to [52, Chapter 1] for an introduction to Itô integrals. Next, we introduce the notation of stochastic differentials for so–called Itô processes.

Definition 1.1.3. *Let X be a continuous adapted stochastic process and let $I = [0, T]$. Then X is said to be a d–dimensional Itô process on the interval I if there exists $b \in L^1(I)^d$ and $g \in L^2(I)^{d\times d}$ such that for every $t \in I$*

$$X(t) = X(0) + \int_0^t b(s)\,ds + \int_0^t g(s)\,dW(s).$$

In this case, we introduce the equivalent notation using stochastic differentials

$$dX(t) = b(t)\,dt + g(t)\,dW(t), \quad t \in I.$$

We have the following fundamental relation between derivative and integral.

Lemma 1.1.4. *(Itô's formula)*
Let X be a d–dimensional Itô process, $V \in C^{1,2}(I \times \mathbb{R}^d)$, $b \in L^1(I)^d$ and $g \in L^2(I)^{d\times d}$ with stochastic differential

$$dX(t) = b(t)\,dt + g(t)\,dW(t).$$

Then $\{V(t, X(t))\}_{t\in I}$ is also a d–dimensional Itô process with stochastic differential

$$\begin{aligned} dV(t, X(t)) =& \Big(\partial_t V(t, X(t)) + \nabla V(t, X(t))b(t) + \frac{1}{2}\operatorname{trace}\big(g(t)^\top \nabla_x^2 V(t, X(t))g(t)\big)\Big)\,dt \\ &+ \nabla V(t, X(t))g(t)\,dW(t), \end{aligned}$$

or equivalently, for all $0 \le t_0 < t \le T$,

$$\begin{aligned} V(t, X(t)) = V(t_0, X(t_0)) &+ \int_{t_0}^t \Big(\partial_t V(s, X(s)) + \nabla V(s, X(s))^\top b(s)\Big)\,ds \\ &+ \int_{t_0}^t \nabla V(s, X(s))^\top g(s)\,dW(s) + \frac{1}{2}\int_{t_0}^t \operatorname{trace}\big(g(s)^\top \nabla^2 V(s, X(s))g(s)\big)\,ds, \end{aligned}$$

where $\nabla = \nabla_x = (\partial_{x_1}, \ldots, \partial_{x_d})^\top$ denotes the gradient w.r.t. the second argument of V.

With the definition of Itô integrals and Itô processes, we are ready to study stochastic differential equations.

Definition 1.1.5. *Let $b : I \times \mathbb{R}^d \to \mathbb{R}^d$, $g : I \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be Borel–measurable. Let $t_0 \in I$ and x_0 be an integrable random variable with $\mathbb{E}[|x_0|^2] < \infty$. We call the stochastic process $X = \{X.(t)\}_{t\in I}$ a solution to the SDE*

$$dX(t) = b(t, X(t))\,dt + g(t, X(t))\,dW(t), \quad X(t_0) = x_0 \tag{1.2}$$

if the following holds:

i) X is continuous and adapted,

ii) $t \mapsto b(t, X(t)) \in L^1(I)^d$ and $t \mapsto g(t, X(t)) \in L^2(I)^{d\times d}$,

iii) for all $t \in I$ it holds that

$$X(t) = x_0 + \int_{t_0}^t b(s, X(s))\,ds + \int_{t_0}^t g(s, X(s))\,dW(s) \quad P\text{–almost surely.}$$

With the use of semicolons, we can include the dependencies of the initial value (t_0, x_0) to the notation of a solution $t \mapsto X(t) = X(t; t_0, x_0)$. Analogously to ordinary differential equations, the mapping $X : [0,T] \times [0,T] \times \Omega \to L^1(\Omega)^d$ is referred to as flow of (1.2).
It is our aim to show that given a solution X to the SDE (1.2), its Lebesgue density function solves a partial differential equation. More precisely, for every $t \in I$, we know that $X(t)$ is a random variable and therefore yields a (time–dependent) PDF $f(t,\cdot) : \mathbb{R}^d \to [0,\infty]$. Hence, we will show that this PDF, considered as a function on $[0,T] \times \mathbb{R}^d$, is a solution of the Fokker–Planck equation under suitable regularity assumptions.
For this purpose, we need to establish the principle of stochastic characteristics, that is, we show that solutions to a related adjoint problem are constant along solutions of (1.2). We fix the time interval to $I = [0,T]$ and consider the elliptic operator

$$L^* := \sum_{i=1}^d b_i(t,x)\partial_{x_i} + \frac{1}{2}\sum_{i,j=1}^d a_{ij}(t,x)\partial^2_{x_i x_j}. \tag{1.3}$$

We impose the following regularity conditions:

(S1) The functions a_{ij}, b_i are bounded on $[0,T] \times \mathbb{R}^d$ and uniformly Lipschitz continuous in (t,x) on any compact subset of $[0,T] \times \mathbb{R}^d$. The functions a_{ij} are Hölder continuous in x, uniformly with respect to (t,x) on $[0,T] \times \mathbb{R}^d$.

(S2) The functions a_{ij} are elliptic in the sense that there exists $\theta > 0$ such that

$$\sum_{i,j=1}^d \xi_i\, a_{ij}(t,x)\, \xi_j \geq \theta|\xi|^2, \quad (t,x) \in [0,T] \times \mathbb{R}^d,\ \xi \in \mathbb{R}^d.$$

Furthermore, $a_{ij} = a_{ji}$ and we define $g : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as its square root, that is, $a_{ij}(t,x)$ is the ij–th entry of the matrix product $g(t,x)^\top g(t,x)$.

(S3) The function ϕ, introduced below, is continuous on $\mathbb{R}^d$ and satisfies the following growth condition: there exists $\alpha, C > 0$ such that

$$|\phi(x)| \leq C(1 + |x|^\alpha), \quad x \in \mathbb{R}^d.$$

Lemma 1.1.6. *Let $q \in C^{1,2}(I \times \mathbb{R}^d)$ be a solution of the Kolmogorov backward equation*

$$\partial_t q = L^* q \quad \text{on } [0,T] \times \mathbb{R}^d, \qquad q(T) = \phi \quad \text{on } \mathbb{R}^d. \tag{1.4}$$

Let X be the flow of the corresponding SDE, that is, for $(t,x) \in [0,T[\times\mathbb{R}^d$, the map $s \mapsto X(s;t,x)$ solves

$$dX(s;t,x) = b(s, X(s;t,x))\, ds + g(s, X(s;t,x))\, dW(s), \quad s \in [t,T[\,, \tag{1.5}$$

$$X(t;t,x) = x. \tag{1.6}$$

Then for $(t,x) \in [0,T] \times \mathbb{R}^d$ it holds that

$$q(t,x) = \mathbb{E}\big[\phi(X(T;t,x))\big].$$

Proof. Let $(t,x) \in [0,T[\times\mathbb{R}^d$ be arbitrary but fix and let us write $X(s) = X(s;t,x)$. By an application of Itô's formula (with $V = q$), we obtain for $s \in]t,T]$

$$q(s,X(s)) = q(t,X(t)) + \int_t^s \Big(\partial_t q(\tau,X(\tau)) + \sum_{i=1}^d b_i(\tau)\partial_{x_i} q(\tau,X(\tau))\Big)\, d\tau$$
$$+ \int_t^s g(\tau)\nabla q(\tau,X(\tau))\, dW(\tau) + \frac{1}{2}\int_t^s \sum_{i,j=1}^d a_{ij}(\tau,x)\partial^2_{x_i x_j} q(\tau,X(\tau))\, d\tau.$$

Since q solves (1.4) and $X(t) = X(t;t,x) = x$, this equation can be simplified to

$$q(s, X(s)) = q(t, x) + \int_t^s g(\tau)\nabla q(\tau, X(\tau))\, dW(\tau).$$

Next, we use the fact that the expected value of any Itô integral is zero. Thus, for $s = T$ with $q(T) = \phi$ on $\mathbb{R}^d$, we obtain

$$\mathbb{E}\big[q(s, X(s;t,x))\big] = q(t,x), \quad s \in]t,T]. \tag{1.7}$$

The choice $s = T$ and the fact that $q(T) = \phi$ on $\mathbb{R}^d$ concludes the proof. □

Remark: Notice that (1.7) implies that q is constant along stochastic characteristics in the following sense

$$\frac{d}{ds}\mathbb{E}\big[q(s, X(s;t,x))\big] = 0, \quad (t,x) \in [0,T] \times \mathbb{R}^d.$$

Furthermore, assuming that we can interchange $\frac{d}{ds}$ and $\mathbb{E}[\cdot]$, we can rewrite $\partial_t q = L^* q$ to

$$\partial_t\Big(\mathbb{E}\big[\phi(X(T;t,x))\big]\Big) = \mathbb{E}\big[L^*\phi(X(T;t,x)\big], \quad (t,x) \in [0,T] \times \mathbb{R}^d.$$

Due to the property

$$\mathbb{E}\big[\varphi(X(s;t,x))\big] = \varphi(x) + \int_t^s \mathbb{E}\big[L^*\varphi(X(\tau;t,x))\big]\, d\tau, \qquad \varphi \in C_c^2(\mathbb{R}^d), \tag{1.8}$$

which makes the connection between L^* and X clear in a different way, L^* is said to be the infinitesimal generator of the stochastic process X. Equation (1.8), in a far more general setting, is also known as Dynkin's formula. In this context, the Fokker–Planck equation, also called Kolmogorov forward equation, is given as the L^2–adjoint. Hence, the elliptic operator from the forward equation L is given as the unique operator that satisfies

$$\int_{\mathbb{R}^d} \varphi(x)(Lf)(x)\, dx = \int_{\mathbb{R}^d} (L^*\varphi)(x) f(x)\, dx \tag{1.9}$$

for all test functions $f, \varphi \in C_c^2(\mathbb{R}^d)$. From (1.3) we consequently obtain

$$Lf(t,x) = \frac{1}{2}\sum_{i,j=1}^d \partial^2_{x_i x_j}(a_{ij}(t,x) f(t,x)) - \sum_{i=1}^d \partial_{x_i}(b_i(t,x) f(t,x)). \tag{1.10}$$

Finally, we can state the main theorem of this section.

Theorem 1.1.7. *Let x_0 be a random variable with PDF $f_0 \in C^2(\mathbb{R}^d)$. Let X be the solution of the corresponding SDE, that is, $I \ni t \mapsto X(t)$ solves*

$$dX(t) = b(t, X(t))\, dt + g(t, X(t))\, dW(t), \quad X(0) = x_0. \tag{1.11}$$

Furthermore, let X have a PDF $f \in C^{1,2}(I \times \mathbb{R}^d)$.
Then, f is the solution of the Fokker–Planck problem

$$\partial_t f = Lf \quad \text{on } I \times \mathbb{R}^d, \qquad f(0) = f_0 \quad \text{on } \mathbb{R}^d. \tag{1.12}$$

Proof. Let $\varphi \in C_c^\infty(\mathbb{R}^d)$ be an arbitrary test function. First, we consider the Kolmogorov backward problem. Since L^* is the infinitesimal generator of X, we obtain

$$\mathbb{E}[\varphi(X(t)] = \mathbb{E}[\varphi(x_0)] + \int_0^t \mathbb{E}[L^*\varphi(X(s))]\, ds$$

On the other hand, since f is the PDF of X, by Lemma 1.1.1 it holds that

$$\mathbb{E}(\varphi[X(t)]) = \int_{\mathbb{R}^d} \varphi(x)f(t,x)\,dx, \quad \mathbb{E}[\varphi(x_0)] = \int_{\mathbb{R}^d} \varphi(x)f_0(x)\,dx,$$
$$\int_0^t \mathbb{E}\big[L^*\varphi(X(s))\big]\,ds = \int_0^t \int_{\mathbb{R}^d} L^*\varphi(x)f(s,x)\,dx\,ds.$$

We combine both results and since L^* and L are adjoint to one–another, we conclude that

$$\begin{aligned}\int_{\mathbb{R}^d} \Big(\varphi(x)f(t,x) - \varphi(x)f_0(x)\Big)\,dx &= \int_0^t \int_{\mathbb{R}^d} L^*\varphi(x)f(s,x)\,dx\,ds \\ &= \int_0^t \int_{\mathbb{R}^d} \varphi(x)Lf(s,x)\,dx\,ds.\end{aligned}$$

By the continuity of Lf and f, and since φ was arbitrary, it follows that for all $(t,x) \in I \times \mathbb{R}^d$

$$f(t,x) = f_0(x) + \int_0^t (Lf)(s,x)\,ds.$$

Lastly, taking ∂_t on both sides and using the fundamental theorem of calculus, we have proven that

$$\partial_t f(t,x) = Lf(t,x), \quad (t,x) \in I \times \mathbb{R}^d.$$

□

We close this section by a verification of Theorem 1.1.7 for the trivial case where the SDE is just a Brownian motion with no force term:

Example 1.1.8. *On $\mathbb{R}^2$ we want to derive the PDF of a Brownian motion of a large particle through suspended medium with Theorem 1.1.7. For that purpose, we consider the following SDE*

$$dX(t) = dW(t), \quad t \in [0,T], \qquad X(0) = x_0.$$

It has the unique solution $X(t) = W(t)+x_0$. In this example, we assume that the particle is with certainty at the origin at $t = 0$, and hence, the PDF of the random variable x_0 is a delta distribution δ_0 at zero. A ccording to Theorem 1.1.7, the corresponding PDE is the heat equation, delta distributed at $t = 0$

$$\partial_t f - \frac{1}{2}\Delta f = 0 \quad \textit{on }]0,T] \times \mathbb{R}^2, \qquad f(0) = \delta_0.$$

It is well known that for $t > 0$, $x \in \mathbb{R}^2$, it holds that

$$f(t,x) := \frac{1}{2\pi t} \exp\left(-\frac{|x|^2}{2t}\right)$$

is a classical solution, and $f(t,\cdot) \to \delta_0$ in the distributional sense as $t \to 0$. On the other hand, f is a Gaussian normal distribution with mean 0 and variance t. According to Definition 1.1.2, the Wiener process, and hence $X(t)$ for $t > 0$, is also normal distributed with the same mean and variance. Due to uniqueness, the PDF of $X(\cdot)$ and the function f have to coincide. This concludes the example.

In this sense, one may say that a Brownian motion can be modelled with the heat equation, and we have successfully verified Theorem 1.1.7.

We remark that we have not investigated the case of boundary conditions, that is, having constraints on the motion of the particle $t \mapsto X(t)$. A typical example is that the particle cannot leave a certain bounded domain $\Omega \subset \mathbb{R}^d$, and therefore, $P(X(t) \notin \Omega) = 0$, i.e., the probability of finding this particle outside the domain is zero. For this purpose, one has to introduce $\tau := \inf\{t \geq 0 \mid X(\tau) \notin \Omega\}$ which is the so–called stopping time or first exit time from Ω for the stochastic process X. This stopping time has to be included into the definition of the adjoint problem from Lemma 1.1.6 and the flow of the corresponding SDE. We omit the details due to the fact that the computation become very lengthy, and we refer the reader to [50, Section 3]. The inclusion of a stopping time yields now Neumann boundary conditions for the backward problem (1.4)

$$\nabla q(t,x)\cdot \hat{n}(x) = 0 \quad \text{on } [0,T]\times\partial\Omega,$$

and so–called reflecting boundary conditions for the Kolmogorov forward problem (1.12)

$$\sum_{i,j=1}^{d} \big(\partial_{x_i}\big(a_{ij}(t,x)\,p(t,x)\big) - \big(b_j(t,x)\,p(t,x)\big)\hat{n}_j(x)\big) = 0 \quad \text{on } [0,T]\times\partial\Omega,$$

where $\hat{n}$ denotes the outward pointing unit normal at each point on the boundary $\partial\Omega$.

1.2 The formulation of objective functionals for particles and their probability distribution functions

I can calculate the motion of heavenly bodies but not the madness of people.

Isaac Newton, 1642 – 1727

In traditional optimal control problems, the goal is to find a control policy that minimizes a certain cost function for a given deterministic system. However, in many real–world scenarios, uncertainties play a significant role, and stochastic optimal control addresses this by considering systems with random variables. To illustrate this, let us once again consider a large particle in a suspended medium of smaller particles, introduced in Section 1.1. Now, we assume that the large particle is driven by a controlled force field with the aim to follow a certain path and to reach a terminal position at time T. The force field

$$b[u](t,x) = F(t,x) + u(t,x)$$

now includes a control function $u = u(t,x)$, that depends on the time t and position x and is an element of a suitable set of admissible controls U_{ad}. The function F denotes a given, exterior force field. Notice that in this control problem, the evolution $t \mapsto X(t)$ is a stochastic process and thus putting $X(t)$ into an objective functional J results in a random variable. For this reason, in the framework of stochastic optimal control, the following averaged objective is analyzed

$$J(X,u) := \mathbb{E}\left[\int_0^T \mathcal{R}(t,X(t),u(t))\,dt + \mathcal{T}(X(T))\right] \tag{1.13}$$

for suitable functions $\mathcal{R}$ and $\mathcal{T}$. Let us investigate, how the stochastic optimal control problem

$$\min_{u\in U_{\mathrm{ad}}} J(X,u) \quad X \text{ subject to} \tag{1.14}$$

$$dX(t) = b[u](t,X(t))\,dt + g(t,X(t))\,dW(t), \quad X(t_0) = x_0 \tag{1.15}$$

can be reformulated in a deterministic framework.

In the previous section, we have built the bridge from investigating stochastic processes X to investigating its distribution function f. More precisely, we have shown that the stochastic processes X, given by the SDE

$$X(t) = x_0 + \int_{t_0}^{t} b(s,X(s))\,ds + \int_{t_0}^{t} g(s,X(s))\,dW(s), \quad t\in[0,T],$$

can be described in a (mathematical) deterministic setting, that is, find its PDF f, given by the following parabolic PDE initial value problem

$$\partial_t f = Lf \quad \text{on } I \times \mathbb{R}^d, \qquad f(0) = f_0 \quad \text{on } \mathbb{R}^d.$$

Furthermore, due to the averaged formulation of the cost functional J, we obtain, after exchanging the integral $\int_0^T$ and the expected value by Fubini and applying Lemma 1.1.1, the following

$$\begin{aligned} J(X,u) &= \mathbb{E}\left[\int_0^T \mathcal{R}(t,X(t),u(t))\,dt + \mathcal{T}(X(T))\right] \\ &= \int_0^T \int_{\mathbb{R}^d} \mathcal{R}(t,x,u(t)) f(t,x)\,dx\,dt + \int_{\mathbb{R}^d} \mathcal{T}(x) f(x,T)\,dx. \end{aligned}$$

This yields a deterministic formulation of the objective, where J is now considered to be a function of f instead of the stochastic process X

$$J(f,u) = \int_0^T \int_{\mathbb{R}^d} \mathcal{R}(t,x,u(t)) f(t,x)\,dx\,dt + \int_{\mathbb{R}^d} \mathcal{T}(x) f(x,T)\,dx.$$

In conclusion, we have proven that under suitable integrability and regularity assumptions on the PDF f and $\mathcal{R}, \mathcal{T}$, the stochastic optimal control problem (1.14) is equivalent to the PDE optimal control problem

$$\min_{u \in U_{\text{ad}}} J(f,u) \quad f \text{ subject to} \tag{1.16}$$

$$\partial_t f = Lf \quad \text{on } I \times \mathbb{R}^d, \qquad f(0) = f_0 \quad \text{on } \mathbb{R}^d. \tag{1.17}$$

We remark that $L = L[u]$, given in (1.10), depends on u. Furthermore, assuming the total force field b takes the form $b(t,x) = F(t,x) + u(t,x)$ for given F, we find that

$$L[u]f = \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j}(a_{ij} f) - \operatorname{div}(F\,f) - \operatorname{div}(u\,f)$$

belongs to the class of so–called bilinear problems in (f,u) due to the last term.
Before we conclude this section, we remark that in many cases, an optimal control problem (1.16)–(1.17) can be written as a minimization problem of the form

$$\min_{u \in U_{\text{ad}}} \hat{J}(u). \tag{1.18}$$

In that case, the PDE constraint (1.17) is built into the definition of the functional $\hat{J}$. Assuming the existence of a well–defined control–to–state map $u \mapsto G(u) = f$, that maps a control u to the (unique) solution f of (1.17), we may introduce the so–called reduced cost functional

$$\hat{J}(u) := J(G(u),u), \quad u \in U_{\text{ad}}.$$

From a theoretical point of view, it is more suitable to study problems of this form rather than constrained minimization problems of the form (1.16)–(1.17). With this in mind, we focus solely on minimization problems of the form (1.18) in the next section.

1.3 Optimization in finite– and infinite–dimensional Banach spaces and the importance of first– and second–order analysis

First– and second–order optimality conditions are important tools for solving minimization problems. In this section, we recall the basic definitions and concepts for addressing infinite–dimensional optimization problems. Furthermore, we present a recent result from Tröltzsch and Casas [21] concerning sufficient second-order optimality conditions, which will be crucial for the analysis of the bilinear optimal control problems under consideration. Before delving deep into this theory, we motivate first– and second-order analysis of optimization problems. Additionally, we highlight, with two examples, the pitfalls that can arise in the transition from finite–dimensional optimization problems to infinite–dimensional ones.

Let us for the moment consider a finite–dimensional minimization problem, that is, we want to find local minima of a smooth function $f : \mathbb{R}^n \to \mathbb{R}$. It is well–known that a necessary condition for local minima $\bar{x}$ of f is $\nabla f(\bar{x}) = 0$. In other words, it is the aim of a first–order analysis to characterize the set of critical points $\{x \in \mathbb{R}^n \mid \nabla f(x) = 0\}$, in which we can find all local minima of f. However, if the minimization problem is non–convex, not all critical points are necessarily local minima and therefore, we have to invoke second–order conditions to find the local minima among the set of critical points. In the finite–dimensional framework, the positive definiteness of the Hessian $\nabla^2 f$ at a critical point $\bar{x}$, that is, $y^\top \nabla^2 f(\bar{x}) y > 0$ for all $y \in \mathbb{R}^n \backslash \{0\}$, is a sufficient optimality condition. Moreover, this condition is equivalent to the positivity of the smallest eigenvalue λ of the corresponding symmetric matrix, i.e., $y^\top \nabla^2 f(\bar{x}) y \geq \lambda |y|^2$ for all $y \in \mathbb{R}^n$.

Now let us consider the infinite–dimensional case $J : U \to \mathbb{R}$, where U is a Banach space and J is differentiable. Typical examples of U that we consider throughout this thesis are the Lebesgue spaces L^2 and L^∞ and the Sobolev space H^1 on a bounded domain. At first, one needs to clarify the concept of differentiation of J on a Banach space U, which leads us to the notion of Fréchet derivatives, given below. Next, it turns out that the necessary first–order optimality conditions for the finite– and infinite–dimensional case are remarkably similar, that is, once again all local minima $\bar{u} \in U$ of J are found in the set of critical points $\{u \in U \mid J'(u) = 0\}$. When it comes to second–order conditions, however, there are significant discrepancies, which are discussed next.

First, the positive definiteness $J''(\bar{u})(v, v) > 0$ is in general not equivalent to the coercivity $J''(\bar{u})(v, v) \geq \Lambda \|v\|_U^2$ for some constant $\Lambda > 0$. It is well–known that on the one hand, coercivity at a critical point $\bar{u}$ – in the correct setting – implies that $\bar{u}$ is a unique local minimum of J; the proof is essentially the same as in the finite–dimensional case. On the other hand, positive definiteness is generally not a sufficient condition for optimality, as demonstrated in the following example.

Example 1.3.1. *Let $U = L^\infty(0, 1)$ and*

$$J(u) := \int_0^1 u(t)^2 \, (t - u(t)) \; dt. \tag{1.19}$$

The zero–function $\bar{u}(t) = 0$ for $t \in]0, 1[$ is a critical point of J and fulfills the positive definiteness since for all $v \in L^\infty(0, 1) \backslash \{0\}$,

$$\begin{aligned} J'(\bar{u})v &= \int_0^1 \bar{u}(t) \, (2t - 3\bar{u}(t)) \, v(t) \, dt = 0, \\ J''(\bar{u})(v, v) &= \int_0^1 (2t - 6\bar{u}(t)) \, v(t)^2 \, dt = \int_0^1 2t \, v(t)^2 \, dt > 0. \end{aligned}$$

Therefore, if the second–order theory from the finite–dimensional setting was correct, one could conclude that $\bar{u}$ is a local minimizer of J in the $L^\infty(0, 1)$–norm. However, this is not true, and in order to disprove the claim, let us consider the following sequence for $n \in \mathbb{N}$

$$u_n(t) := \begin{cases} 2t & \text{for } t \in]0, 1/n[, \\ 0 & \text{else.} \end{cases}$$

Consequently, $J(u_n) = -1/n^4 < 0 = J(\bar{u})$ and $\|u_n - \bar{u}\|_{L^\infty(0,1)} = 2/n$. Hence, $\bar{u}$ is not a local minimum of J w.r.t. the L^∞–norm. Notice that $J''(\bar{u})$ is also not coercive w.r.t. the L^∞–norm, since there exists no constant $\Lambda > 0$ such that

$$\int_0^1 2t\, v(t)^2\, dt \geq \Lambda \|v\|^2_{L^\infty(0,1)}, \quad \textit{for all } v \in L^\infty(0,1).$$

The next example from [64] introduces the so called two–norm discrepancy. The example makes use of the fact that two norms on an infinite–dimensional Banach space are in general not equivalent. Consequently, if we speak about differentiability and coercivity of J or local uniqueness of minimizers, we need to be precise and consistent which norm we use. Obviously, this situation cannot arise in finite–dimensional problems since all norms are equivalent, however, it is well–known to appear in optimal control problems.

Example 1.3.2. *Define for $u \in L^2(0,1)$ the functional*

$$J(u) := -\int_0^1 \cos\big(u(t)\big)\, dt \tag{1.20}$$

and notice that the zero–function $\bar{u}(t) := 0$ is a global minimizer. Furthermore, $\bar{u}$ satisfies $J'(\bar{u})v = 0$ and is coercive w.r.t. the L^2–norm

$$J''(\bar{u})(v,v) = \int_0^1 \cos(0) v(t)^2\, dt = \|v\|^2_{L^2(0,1)}. \tag{1.21}$$

However, $\bar{u}$ is not a locally unique minimum in the L^2–norm, that is, there are infinitely many different global minimizers of J in any L^2–neighborhood of $\bar{u}$. This can be seen by defining for $0 < \varepsilon < 1$ the function

$$u_\varepsilon(t) := \begin{cases} 0 & \textit{for } t \in]0, \varepsilon[\\ 2\pi & \textit{for } t \in [\varepsilon, 1[, \end{cases}$$

and observing that $\|u_\varepsilon - \bar{u}\|_{L^2(0,1)} = 2\pi\sqrt{\varepsilon}$.

So, despite having coercivity around the local minimum, it is not isolated nor strict. Let us analyze what went wrong. Even though we have used the same norm, $L^2(0,1)$, for the formulation of coercivity and uniqueness, strict local optimality is not obtained in this example. The problem is hidden in the differentiability property of J, more precisely, J is not twice continuously Fréchet differentiable in the space $L^2(0,1)$, cf. [64]. Consequently, we cannot make a statement about coercivity in the L^2–norm. However, we can easily verify with the following definition that J is differentiable in the space $L^\infty(0,T)$, and $J''(\bar{u})$ from (1.21) is the correct L^∞–derivative:

Definition 1.3.3. *Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be two normed spaces, let $M \subset X$, $x \in M$ and $F : M \to Y$. The function $F'(x) : X \to Y$ is said to be the Fréchet derivative of F at x if $F'(x)$ is a linear and bounded operator from X to Y, i.e. $F'(x) \in \mathrm{Lin}(X,Y)$, and*

$$\frac{\|F(x+h) - F(x) - F'(x)h\|_Y}{\|h\|_X} \longrightarrow 0, \quad \textit{as } \|h\|_X \to 0.$$

If this holds for every $x \in M$ such F is called Fréchet differentiable on M from $(X, \|\cdot\|_X)$ to $(Y, \|\cdot\|_Y)$ or is said to be of class C^1 if the spaces and norms are clear from the context. Furthermore, such F is called twice Fréchet differentiable on M from $(X, \|\cdot\|_X)$ to $(Y, \|\cdot\|_Y)$ or of class C^2 if $F' : M \to \mathrm{Lin}(X,Y)$ is of class C^1 on M. We write $F''(x)(a,b) = F''(x)(a)(b)$ and notice that $F'' : M \to \mathrm{Lin}(X; \mathrm{Lin}(X,Y))$, that is, $F''(x) \in \mathrm{Bilin}(X,X)$ is a continuous bilinear mapping for any $x \in M$.

Let us remark that Fréchet differentiability may depend on the specific norms used for the spaces X and Y. Let $\|\cdot\|_{X_s}$, $\|\cdot\|_{X_w}$ and $\|\cdot\|_{Y_s}$, $\|\cdot\|_{Y_w}$ denote two different norms on X and Y, respectively. Let the subscript "s" denote the stronger norm and let "w" denote the weaker norm, that is, there exists some $C > 0$ such that for all $x \in X, y \in Y$

$$\|x\|_{X_w} \leq C\|x\|_{X_s} \quad \text{and} \quad \|y\|_{Y_w} \leq C\|y\|_{Y_s}.$$

Then, it is immediately apparent from the definition that Fréchet differentiability of F from $(X, \|\cdot\|_{X_w})$ to $(Y, \|\cdot\|_{Y_s})$ implies the Fréchet differentiability from $(X, \|\cdot\|_{X_s})$ to $(Y, \|\cdot\|_{Y_w})$, however, the converse is in general wrong. This is what happens in Example 1.3.2; notice that $X_w = L^2(0,1)$ is a weaker norm than $X_s = L^\infty(0,1)$. Therefore, even though $J : L^2(0,1) \to \mathbb{R}$ is well–defined, twice Fréchet differentiable from $L^\infty(0,1)$ to $\mathbb{R}$ and the formula for $J''(u)$ makes sense for $L^2(0,1)$ functions, J is not twice Fréchet differentiable on $L^2(0,1)$. Therefore, J cannot be coercive w.r.t. the L^2–norm, and we cannot apply the classical theory on second–order conditions.

Can we fix this issue by switching everywhere from the L^2–norm to the L^∞–norm? Unfortunately not since $J''(u)$ is not coercive w.r.t. the L^∞–norm. This phenomenon often arises in optimal control problems, where the reduced cost functional is twice Fréchet differentiable only in a stronger norm, e.g., L^∞ but coercivity holds only for a weaker norm, such as L^2.

This motivates the following essential theorem from [21] that allows to consider an optimization problem with two different norms.

Let $(U_2, \|\cdot\|_2)$ be a Hilbert space and $(U_\infty, \|\cdot\|_\infty)$ be a Banach space with continuous embedding $U_\infty \subset U_2$. Let $\emptyset \neq U_{\text{ad}} \subset U_\infty$ be convex and let $A \subset U_\infty$ be an open set covering U_{ad}. The objective reads $J : A \to \mathbb{R}$, and we consider the minimization problem

$$\min_{u \in U_{\text{ad}}} J(u). \tag{1.22}$$

For $\varepsilon > 0$, $j \in \{2, \infty\}$ and $w \in U_j$, we recall the following notation for the open ball around w

$$B_\varepsilon(w; U_j) = \{u \in U_j \mid \|w - u\|_j < \varepsilon\}.$$

Definition 1.3.4. *We say that $\bar{u}$ is a local solution of* (1.22) *or a local minimizer of J in U_∞, if there exists some $\varepsilon > 0$ such that $J(\bar{u}) \leq J(u)$ holds for all $u \in U_{\text{ad}} \cap B_\varepsilon(\bar{u}; U_\infty)$. If $J(\bar{u}) < J(u)$ holds for this set with $u \neq \bar{u}$, we say that $\bar{u}$ is a strict minimizer in U_∞ and locally unique in U_∞.*

Notice that since U_2 is a weaker norm, every local minimizer in U_2 is also a local minimizer in U_∞. The next theorem gives a necessary first–order condition; the proof is essentially the same as in the finite–dimensional case.

Theorem 1.3.5. *Let $\bar{u}$ be a local solution of* (1.22) *and let J be Fréchet differentiable in $\bar{u}$, both in the sense of U_∞. Then,*

$$J'(\bar{u})(u - \bar{u}) \geq 0, \quad u \in U_{\text{ad}}.$$

Next, we specify the conditions on the minimization problem (1.22) that involves a second–order analysis. Let us fix $\bar{u} \in U_{\text{ad}}$.

(C1) $J : A \to \mathbb{R}$ is of class C^2 from $(A, \|\cdot\|)$ to $\mathbb{R}$ and for every $u \in U_{\text{ad}}$ there exists continuous extensions

$$J'(u) \in \text{Lin}\,(U_2)\,, \quad J''(u) \in \text{Bilin}\,(U_2 \times U_2)\,. \tag{C1}$$

(C2) There exists $\Lambda > 0$ such that for all sequences $(u_n)_{n\in\mathbb{N}} \subset U_{\text{ad}}$ and $(v_n)_{n\in\mathbb{N}} \subset U_2$ with $u_n \to \bar{u}$ strongly in U_2 and $v_n \rightharpoonup v$ weakly in U_2:

$$J'(\bar{u})v = \lim_{n\to\infty} J'(u_n)v_n, \tag{C2.1}$$

$$J''(\bar{u})(v,v) \leq \liminf_{n\to\infty} J''(u_n)(v_n,v_n), \tag{C2.2}$$

$$\text{and if } v = 0, \text{ then } \quad \Lambda \liminf_{n\to\infty} \|v_n\|_2^2 \leq \liminf_{n\to\infty} J''(u_n)(v_n,v_n). \tag{C2.3}$$

In addition to these conditions, we require the following standard first– and second–order assumptions on $\bar{u}$

$$J'(\bar{u})(u-\bar{u}) \geq 0, \quad u \in U_{\text{ad}}, \tag{A1}$$

$$J''(\bar{u})(v,v) > 0, \qquad v \in C_{\bar{u}}\backslash\{0\}. \tag{A2}$$

with the sets

$$\begin{aligned} S_{\bar{u}} &:= \{\lambda(u-\bar{u}) \,:\, \lambda > 0 \text{ and } u \in U_{\text{ad}}\}, &&\text{(cone of feasible directions)}\\ C_{\bar{u}} &:= \overline{S_{\bar{u}}}^{U_2} \cap \{v \in U_2 \,:\, J'(\bar{u})v = 0\}, &&\text{(critical cone).} \end{aligned}$$

Due to Theorem 1.3.5, assumption (A1) is called first–order necessary condition (FONC), and functions $\bar{u}$ satisfying (A1) are critical points of J.
Although proving these properties of J turns out to be very challenging, it is certainly rewarding in terms of the statements about quadratic growth conditions, local uniqueness and coercivity of minimizers.

Theorem 1.3.6. *Let $\bar{u} \in U_{\text{ad}}$ and J satisfy* (A1)–(A2) *and* (C1)–(C2.3)*, respectively. Then there exists $\varepsilon, \delta, \nu, \tau > 0$ such that the following holds.*

a) For all $u \in U_{\text{ad}} \cap B_\varepsilon(\bar{u}; U_2)$, it holds that

$$J(\bar{u}) + \frac{\delta}{2}\|u-\bar{u}\|_2^2 \leq J(u).$$

b) For all critical points u^* with $u^* \in U_{\text{ad}} \cap B_\varepsilon(\bar{u}; U_2)$, it holds that

$$\bar{u} = u^*.$$

c) For all $u \in U_{\text{ad}} \cap B_\varepsilon(\bar{u}; U_2)$ and all $v \in E_{\bar{u}}^\tau$, it holds that

$$J''(u)(v,v) \geq \frac{\nu}{2}\|v\|_2^2,$$

where $E_{\bar{u}}^\tau := \{v \in \overline{S_{\bar{u}}}^{U_2} \,:\, |J'(\bar{u})v| \leq \tau\|v\|_2\}$.

The proofs can be found in [21]. Before we conclude this section, let us discuss the meaning and application of each statement. Part a) is called the quadratic growth condition and implies that $\bar{u}$ is a strict local minimizer. Hence, in this setting, (A2) may be referred to as sufficient second–order condition (SSC). Furthermore, this estimate can be used as a starting point in order to derive accuracy or stability estimates. We remark that the numerical verification of the SSC condition is generally challenging, and we refer to [58] for the case of a semilinear elliptic optimal control problem. Assertion b) states that the local minima $\bar{u}$ is isolated, that is, there are no other critical points u^* – with possibly different value $J(u^*)$ – close to it. This statement is of high relevance in the numerical calculation of local minima $\bar{u}$, since in standard methods the zeros of J' are searched for, close to the presumed minimum. Thus, if there are infinitely many critical points close to $\bar{u}$, there is no chance for this procedure to be successful. There cannot be any guarantee that the critical point one computes is the desired local minima $\bar{u}$, no matter how close to $\bar{u}$ one starts. Lastly, part c) is called a local coercivity condition on $\hat{J}$ on the extended cone E_u^τ; notice that $C_{\bar{u}} = E_{\bar{u}}^0 \subset E_{\bar{u}}^\tau$. Coercivity is essential for the numerical analysis. In view of a Taylor expansion of $\hat{J}$ around $\bar{u}$, we may say that coercivity implies local convexity of $\hat{J}$.

1.4 Auxiliary results

Throughout this thesis, a few arguments or techniques will appear more frequently. Therefore, we have collected important assertions in the following Lemma.

Lemma 1.4.1. *(An Application of Egorov's theorem)*
Let $M \subset \mathbb{R}^d$ be open and bounded. Let $(f_k)_{k\in\mathbb{N}} \subset L^\infty(M)$ be non–negative a.e. on M with $f_k \to f$ in $L^1(M)$ and $\|f_k\|_{L^\infty(M)} < C$ for all $k \in \mathbb{N}$. Furthermore, let $(v_k)_{k\in\mathbb{N}} \subset L^2(M)$ with $v_k \rightharpoonup v$ in $L^2(M)$. Then, it holds that

$$\int_M |v(x)|^2 f(x)\, dx \le \liminf_{k\to\infty} \int_M |v_k(x)|^2 f_k(x)\, dx.$$

Proof. A proof can be found in [21, Lemma 3.5]. □

Lemma 1.4.2. *(Mazur's lemma)*
Let $1 < q < \infty$, let $M \subset \mathbb{R}^d$ be open and measurable, let $(g_j)_{j\in\mathbb{N}} \subset L^q(M)$ with $g_j \rightharpoonup G$ in $L^q(M)$. Then, there exists a convex combination G_j of the functions $g_1, \dots, g_j$ such that the sequence $(G_j)_{j\in\mathbb{N}}$ converges strongly to G in $L^q(M)$.

Proof. We refer the reader to [49, Theorem 2.13] for a proof. □

We recall that some f is a convex combination of $g_1, \dots, g_j$ if there exists $\lambda^i \in [0,1]$, $i = 1, \dots, j$, such that $\sum_{i=1}^j \lambda_i = 1$ and $f = \sum_{i=1}^j \lambda^i g_i$.

Lemma 1.4.3. *(Extraction of a subsequence)*
Let $(X, \|\cdot\|_X)$ be a normed space and $(x_k)_{k\in\mathbb{N}} \subset X$. Then, (i) and (ii) are equivalent:

(i) There exists $x \in X$ such that

$$x_k \to x \text{ in } X \text{ as } k \to \infty.$$

(ii) There exists $x \in X$ such that every subsequence $\left(x_{k(n)}\right)_{n\in\mathbb{N}}$ of $(x_k)_{k\in\mathbb{N}}$ has a convergent sub–subsequence $\left(x_{k(n(j))}\right)_{j\in\mathbb{N}}$ such that

$$x_{k(n(j))} \to x \quad \text{in } X \quad \text{as } j \to \infty.$$

We remark that in (ii), the limit x may not depend on the selection of subsequence $k(n)$.

Proof. The implication from (i) to (ii) is trivial. Now assume that the implication from (ii) to (i) is not true. Hence, we find $\epsilon > 0$ and a selection of subsequences $n \mapsto k(n)$ such that $|x - x_{k(n)}| \ge \epsilon$ for all $n \in \mathbb{N}$. Consequently, this subsequence $(x_{k(n)})_{n\in\mathbb{N}}$ has no converging sub–subsequence to x, however, this is a contradiction to (ii). □

The following special case of Grönwall's inequality appears multiple times throughout this thesis.

Lemma 1.4.4. *(Grönwall's inequality)*
Let $\alpha \geq 0$ *and* $0 \leq u, \beta \in C([0,T])$ *with*

$$u(t) \leq \alpha + \int_0^t \beta(s)u(s)\,ds, \quad t \in [0,T].$$

Then, u satisfies the explicit inequality

$$u(t) \leq \alpha \exp\left(\int_0^t \beta(s)\,ds\right)), \quad t \in [0,T].$$

Proof. Let $x \in C^1([0,T])$ be the unique solution to the linear initial value problem $x'(t) = \beta(t)x(t)$ with $x(0) = \alpha$. Hence, on the one hand, $x(t) = \alpha + \int_0^t \beta(s)x(s)\,ds$, which implies $u(t) \leq x(t)$ for all $t \in [0,T]$. On the other hand, x is given by $x(t) = \alpha \exp\left(\int_0^t \beta(s)\,ds\right)$. This proves the claim. □

Lemma 1.4.5. *(Higher regularity of elliptic problems)*
Let $M \subset \mathbb{R}^d$ *be a bounded domain. Let* M *have* C^2*–boundary or let* M *be polygonal and convex. Let* $2 \leq q < \infty$, $f \in L^q(M)$ *and let* $u \in H_0^1(M)$ *be a weak solution to the elliptic problem*

$$-\Delta u = f \quad \text{in } M, \qquad u = 0 \quad \text{in } \partial M.$$

Then, $u \in W^{2,q}(M)$ *and there exists* $C = C(\Omega, q)$ *such that*

$$\|u\|_{W^{2,q}(M)} \leq C\|f\|_{L^q(M)}.$$

We remark that an analogous estimate holds if $\Omega \subset \mathbb{R}$ is an open interval.

Proof. A proof is given in [38, Chapter 4] for the polygonal case, and in [34, Chapter 6] for the case of smooth ∂M. □

Notice that regularity results for the Poisson problem $-\Delta u = f$ cover the regularity of all elliptic problems $-\Delta u + R(u) = f$, where $R(u)$ contains all lower order derivatives of u. To see this, simply apply Lemma 1.4.5 with r.h.s $f - R(u) \in L^2(M)$.
For the following lemma, let $-\infty < a < 0 < b < \infty$ and let $\Omega \subset \mathbb{R}^d$ be a non–empty bounded domain. Let us introduce the set of constrained test functions

$$L_a^b := \{u \in L^\infty(\Omega) \mid a \leq u(x) \leq b \quad \text{f.a.e. } x \in \Omega\}.$$

Lemma 1.4.6. *(Variational inequalities with constraints)*
Let $f \in L^2(\Omega)$ *and* $u \in L_a^b$ *with*

$$\langle f, v - u\rangle_{L^2(\Omega)} \geq 0 \quad \text{for all } v \in L_a^b.$$

Then, for any measurable set $M \subset \Omega$*, it holds that*

$$\begin{cases} f(x) > 0 \quad \text{f.a.e. } x \in M & \Longrightarrow \; u(x) = a \quad \text{f.a.e. } x \in M, \\ f(x) < 0 \quad \text{f.a.e. } x \in M & \Longrightarrow \; u(x) = b \quad \text{f.a.e. } x \in M, \\ a < u(x) < b \quad \text{f.a.e. } x \in M & \Longrightarrow \; f(x) = 0 \quad \text{f.a.e. } x \in M. \end{cases}$$

Proof. Let us start with the first implication. We may assume that, after changing f on a set of measure zero, $M \subset \Omega$ is either the empty set or has positive volume, and f is positive everywhere on M. Next, define for a pointwise defined representant of u

$$v(x) := \begin{cases} a & \text{if } x \in M, \\ u(x) & \text{if } x \in \Omega \backslash M. \end{cases} \tag{1.23}$$

By construction, we have that $v \in L_a^b$ is a valid test function, and hence

$$0 \leq \langle f, v - u \rangle_{L^2(M)} + \langle f, v - u \rangle_{L^2(\Omega \backslash M)}.$$

The last term is zero, since $v = u$ on $\Omega \backslash M$. Furthermore, notice that $f > 0$ on M and $v - u = a - u \leq 0$ on M. Now we show that $u = a$ on M almost everywhere. Assume that this is not true, i.e., there exists a subset M' of M with $\mathrm{vol}(M') > 0$ on which $u \neq a$. Therefore, $v - u < 0$ on M'. This, however, is an immediate contradiction to

$$0 \leq \langle f, v - u \rangle_{L^2(M)}$$

and we have proven that $u = a$ on M.

The second implication can be shown analogously, where obviously a is replaced by b in the definition of v, and the third implication follows from the first and the second one. □

2
The Fokker–Planck equation

Insight must precede application.

Max Planck, 1858 – 1947

Let us consider a drift–diffusion model given by the following SDE

$$dX(t) = B[u](t, X(t))\, dt + \sigma(t, X(t))\, dW(t), \quad X(0) = x_0 \tag{2.1}$$

for $t \in [0, T]$. The function $B[u] : [0, T] \times \Omega \to \mathbb{R}^d$ denotes the drift including a control mechanism u, and $\sigma : [0, T] \times \Omega \to \mathbb{R}^{d\times d}$ represents a diffusion matrix. In Section 1.1, we have shown that the PDF $p : [0, T] \times \Omega \to \mathbb{R}$ of this stochastic process X is given by the FP equation. Since Ω is a bounded domain, we can additionally impose flux–zero boundary condition. Therefore, the FP problem under investigation reads

$$\begin{aligned}
&\partial_t p = \sum_{i,j=1}^{d} \partial^2_{x_i x_j}\big(a_{ij}\, p\big) - \operatorname{div}\big(B[u]\, p\big) && \text{on } \Omega_T, && (2.2)\\
&p(0) = p_0 && \text{on } \Omega, && (2.3)\\
&F \cdot \hat{n} = 0 && \text{on } [0, T] \times \partial\Omega, && (2.4)
\end{aligned}$$

with diffusion $a = \frac{1}{2}\sigma^\top \sigma$ and probability density flux $F = F[p]$, where for $(t, x) \in \Omega_T :=]0, T[\, \times \Omega$

$$F[p]_j(t, x) := \sum_{i=1}^{d} \partial_{x_i}\big(a_{ij}(t, x)\, p(t, x)\big) - \big(B[u]_j(t, x)\, p(t, x)\big), \quad j = 1, \ldots, d.$$

We remark that (2.2) can be rewritten in flux form as follows

$$\partial_t p(t, x) = \operatorname{div} F[p](t, x), \quad (t, x) \in [0, T] \times \Omega. \tag{2.5}$$

We introduce the following assumptions:

(F1) The drift $B[u]$ is of the form

$$B[u](t,x) = M(t,x)u(t,x) + c(t,x), \quad (t,x) \in \Omega_T$$

where $c \in L^\infty(\Omega_T)^d$, $M \in L^\infty(\Omega_T)^{d\times m}$ and $u \in L^\infty(\Omega_T)^m$.

(F2) The coefficients a_{ij} of a, given by the diffusion matrix $a = \frac{1}{2}\sigma^\top\sigma$, enjoy the regularity $W^{1,\infty}(\Omega_T)$. Furthermore, (a_{ij}) is elliptic in the sense that there exists $\theta > 0$ such that

$$\sum_{i,j=1}^{d} \xi_i\, a_{ij}(t,x)\, \xi_j \geq \theta|\xi|^2, \quad (t,x) \in \Omega_T,\ \xi \in \mathbb{R}^d.$$

(F3) The initial distribution p_0 is the PDF of x_0 with regularity $p_0 \in L^\infty(\Omega)$.

Moreover, if higher regularity of solutions to the FP problem is investigated, we may assume the following:

(F4) The initial state enjoys the higher regularity $p_0 \in H^3(\Omega)$.

(F5) $M \in L^\infty(0,T;W^{1,\infty}(\Omega))^{d\times m}$ and for every function u from the set of admissible controls, it holds

$$\big(M(t,x)u(t,x)\big) \cdot \hat{n}(x) = 0, \quad \text{f.a.e. } x \in \partial\Omega, \text{ f.a.e. } t \in [0,T].$$

(F6) It either holds that

(i) $c \in L^\infty(0,T;W^{1,\infty}(\Omega))^d$ and f.a.e. $t \in [0,T]$ it holds that $c(t,\cdot)\cdot\hat{n} = 0$ a.e. on $\partial\Omega$; or

(ii) c has a potential $-V \in C([0,T];W^{2,\infty}(\Omega))$ such that $c = \nabla V$ a.e. on Ω_T .

(F7) The diffusion matrix (a_{ij}) is up to a positive constant – denoted with the same variable $a > 0$ – the identity matrix.

We impose that throughout this chapter, the assumptions (F1)–(F3) hold. For certain statements, we will additionally assume (F4)–(F7) but this will always be mentioned.
Throughout this thesis, we use $C_\mathrm{F} > 0$ as a generic constant that depends on given quantities in (F1)–(F3), i.e., C_F depends continuously on the real valued numbers

$$\|c_i\|_{L^\infty(\Omega_T)}, \|M_{ij}\|_{L^\infty(\Omega_T)}, \|a_{ij}\|_{W^{1,\infty}(\Omega_T)}, \|p_0\|_\infty, \theta, T, \quad i,j = 1,\dots,d \tag{2.6}$$

and on certain embedding constants, depending only on Ω and its dimension d. Furthermore, the generic constant $C_{\mathrm{F}*} > 0$ depends additionally on the quantities of (F4)–(F7), that is

$$\|V\|_{L^\infty W^{2,\infty}} \text{ or } \|c_i\|_{L^\infty W^{1,\infty}} \text{ and } \|M_{ij}\|_{L^\infty W^{1,\infty}}, \|p_0\|_{H^2(\Omega)},, \quad i,j = 1,\dots,d,$$

and $C_u > 0$ denotes a generic constant that depends continuously only on $\|u\|_{L^\infty(\Omega_T)}$ in the case of time–space dependent controls, or on $\|u\|_{L^\infty(0,T)}$ in the case of only time–dependent controls. Lastly, let us recall the abbreviations for norms of Lebesgue spaces for and Sobolev spaces

$$\|\cdot\|_q := \|\cdot\|_{L^q(M)}, \quad q \in [1,\infty], \qquad \|\cdot\|_{H^1} := \|\cdot\|_{H^1(M)},$$

where it will be clear from the context what $M \in \{\Omega,]0,T[\,, \Omega_T\}$ is.

2.1 Existence and uniqueness of weak solutions

We start this section by deriving a weak formulation of the FP problem (2.2)–(2.5). Then, we show existence of weak solutions in the space

$$W(0,T) := H^1(0,T;H^1(\Omega)') \cap L^2(0,T;H^1(\Omega)),$$

and prove that weak solutions satisfy the typical properties of PDFs.
We begin with the derivation of a weak formulation. Let $\psi \in H^1(\Omega)$ be a test function and consider some smooth p that satisfies (2.4)–(2.5). An application of Green's formula yields

$$\begin{aligned}\int_\Omega \partial_t p\, \psi\, dx &= -\int_\Omega \sum_{i,j=1}^d \partial_{x_j}(a_{ij}p)\partial_{x_i}\psi\, dx + \int_\Omega p\, B[u]\cdot\nabla\psi\, dx \\ &\quad + \int_{\partial\Omega} \sum_{i,j=1}^d \partial_{x_j}(a_{ij}\, p)\psi\, \hat{n}_i\, dS(x) - \int_{\partial\Omega} p\, B[u]\cdot\hat{n}\,\psi\, dS(x) \\ &= -\int_\Omega \left(\sum_{i,j=1}^d \partial_{x_j}(a_{ij}p)\partial_{x_i}\psi - p\, B[u]\cdot\nabla\psi\right) dx \\ &= -\int_\Omega F[p]\cdot\nabla\psi\, dx,\end{aligned}$$

a.e.on $[0,T]$. Consequently, we obtain the following bilinear flux–operator

$$\begin{aligned}&\mathcal{F}_t : H^1(\Omega)\times H^1(\Omega) \to \mathbb{R} \quad \text{f.a.e. } t \in]0,T[\,, \\ &\mathcal{F}_t(p,\psi) := \int_\Omega \left[\sum_{i,j=1}^d \partial_{x_i}\big(a_{ij}(t,x)p(x)\big)\,\partial_{x_j}\psi(x) - p(x)\, B[u](t,x)\cdot\nabla\psi(x)\right] dx. \qquad (2.7)\end{aligned}$$

It turns out to be convenient to rewrite $\mathcal{F}$ as follows

$$\mathcal{F}(p,\psi) = \int_\Omega \left(\sum_{i,j=1}^d a_{ij}\partial_{x_j}p\,\partial_{x_i}\psi - p\, b\cdot\nabla\psi\right) dx, \quad p,\psi \in H^1(\Omega), \qquad (2.8)$$

$$\text{with} \quad b_i(t,x) := B[u]_i(t,x) - \sum_{j=1}^d \partial_{x_j}a_{ij}(t,x), \quad (t,x)\in\Omega_T,\ i=1,\dots,d. \qquad (2.9)$$

The well–definedness of $\mathcal{F}$ is shown in Lemma 2.1.2 below. Hence, given the initial distribution p_0 on Ω, we have the following weak solution concept for (2.2)–(2.4).

Definition 2.1.1. *We call $p \in W(0,T)$ a weak solution to the FP problem with flux–zero boundary conditions and initial state p_0 if there exists some null set $N \subset [0,T]$ such that for all $\psi \in H^1(\Omega)$ and all $t \in [0,T]\backslash N$:*

$$\langle \dot{p}(t),\psi\rangle_{H^1(\Omega)'} + \mathcal{F}_t(p(t),\psi) = 0, \qquad p(0) = p_0 \quad \textit{a.e. on } \Omega. \qquad (2.10)$$

We recall that $H^1(\Omega)'$ denotes the dual space of $H^1(\Omega)$ with pivot space $L^2(\Omega)$,

$$H^1(\Omega)' := \{f : H^1(\Omega)\to\mathbb{R} : \langle f,\cdot\rangle_{H'} := \langle f,\cdot\rangle_{H^1(\Omega)'} = f(\cdot) \text{ is linear and continuous}\} \qquad (2.11)$$

and if $f \in L^2(\Omega)$, then $\langle f,\cdot\rangle_{H'} = \langle f,\cdot\rangle_{L^2(\Omega)}$.

In the following, for similar definitions or in similar settings, we will sometimes just write "f.a.e. $t \in [0,T]$" instead of "for all $t \in [0,T]\backslash N$" with the meaning that the set of measure zero in $[0,T]$ is independent of the test function ψ. For later convenience, we define for $W(0,T)$–functions

$$\|\cdot\|_{W(0,T)} := \|\cdot\|_{L^2(0,T;H^1(\Omega))} + \|\partial_t\cdot\|_{L^2(0,T;H^1(\Omega)')}.$$

Furthermore, we recall the continuous embedding

$$W(0,T) \subset C([0,T];L^2(\Omega)), \tag{2.12}$$

which gives meaning to the expression $p_{|t=0} = p(0) \in L^2(\Omega)$. We also remark that the first equation in (2.10) is equivalent to the Bochner space formulation that is used frequently throughout this thesis

$$\dot{p} + \mathcal{F}(p,\cdot) = 0 \quad \text{in } L^2(0,T;H^1(\Omega)'). \tag{2.13}$$

Next, we establish some a–priori bounds.

Lemma 2.1.2. *The flux–operator $\mathcal{F}$ is bounded and weakly coercive, i.e. there exists a null set $N \subset [0,T]$ and constants $C_F, \beta, \gamma > 0$ such that for all $p, \psi \in H^1(\Omega)$, $t \in [0,T]\backslash N$*

$$\begin{aligned} |\mathcal{F}_t(p,\psi)| &\le C_{\mathrm{F}} \|p\|_{H^1} \|\psi\|_{H^1} && \textit{(boundedness)},\\ \beta\|p\|_{H^1}^2 &\le \mathcal{F}_t(p,p) + \gamma\|p\|_2^2 && \textit{(weak coercivity).} \end{aligned}$$

Proof. In order to show boundedness, let $p, \psi \in H^1(\Omega)$ and obtain

$$\begin{aligned} \mathcal{F}_t(p,\psi) &= \int_\Omega \left(\sum_{i,j=1}^d a_{ij}(t,x)\partial_{x_i}p(x)\,\partial_{x_j}\psi(x) + p(x)\,b(t,x)\cdot\nabla\psi(x) \right) dx \\ &\le \left(\sum_{i,j=1}^d \|a_{ij}\|_{L^\infty(\Omega_T)} + \sum_{i=1}^d \|b_i\|_{L^\infty(\Omega_T)} \right) \|p\|_{H^1}\|\psi\|_{H^1}. \end{aligned}$$

For the weak coercivity, we exploit the ellipticity (F2) of a_{ij} with $\xi := \nabla p$. Consequently, we obtain f.a.e. $t \in [0,T]$ that

$$\begin{aligned} \int_\Omega \theta|\nabla p(x)|^2\,dx &\le \int_\Omega \sum_{i,j=1}^d a_{ij}(t,x)\partial_{x_i}p(x)\,\partial_{x_j}p(x)\,dx \\ &\le \mathcal{F}_t(p,p) + \|b\|_{L^\infty(\Omega_T)} \int_\Omega |p(x)||\nabla p(x)|\,dx. \end{aligned}$$

Next, we use the ε–Young's inequality $c_1c_2 \le \varepsilon c_1^2 + c_2^2/(4\varepsilon)$, which holds for any values $c_1, c_2 \in \mathbb{R}$ and $\varepsilon > 0$ arbitrary, and choose $c_1 = |\nabla p|, c_2 = |p|$ and $\varepsilon = \theta/(2\|b\|_\infty)$. Thus, we conclude

$$\frac{\theta}{2}\int_\Omega |\nabla p(x)|^2\,dx \le \mathcal{F}_t(p,p) + \|p\|_2^2\|b\|_\infty/(4\varepsilon), \quad \text{f.a.e. } t \in [0,T]. \tag{2.14}$$

Finally, adding $\frac{\theta}{2}\int_\Omega |p(x)|^2\,dx$ to both sides of (2.14) yields the assertion with constants $\beta := \theta/2$ and $\gamma := \|b\|_\infty^2/(2\theta) + \theta/2$. □

We have the following existence and uniqueness result for the weak FP problem for initial distributions from $L^2(\Omega)$.

Theorem 2.1.3. *For every initial distribution $p_0 \in L^2(\Omega)$, the following holds.*

a) There exists a weak solution $p \in W(0,T)$ of the Fokker–Planck problem with flux–zero boundary conditions and $p(0) = p_0$ in the sense of Definition 2.1.1.

b) There exists some constant $C_F > 0$ independent of p_0 such that

$$\|p\|_{L^\infty(0,T;L^2(\Omega))} + \|p\|_{L^2(0,T;H^1(\Omega))} + \|\partial_t p\|_{L^2(0,T;H^1(\Omega)')} \le C_F C_u \|p_0\|_{L^2(\Omega)}$$

c) p is unique in $C([0,T];L^2(\Omega))$*.*

Proof. The existence of weak solutions can be shown with a standard Galerkin approach. Hence, it is sufficient to show the a–priori estimates stated in b).
Let p denote a weak solution. Due to the continuous embedding (2.12), we can choose one representative $p \in C([0,T];L^2(\Omega))$, which is fixed from now on. This yields the following well–known identities that are used frequently throughout this work

$$p(t) = p(s) + \int_s^t \dot{p}(\tau)\, d\tau \quad \text{and} \quad \|p(t)\|_2^2 = \|p(s)\|_2^2 + 2\int_s^t \langle \dot{p}(\tau), p(\tau)\rangle_{H'}\, d\tau, \quad s,t \in [0,T]. \tag{2.15}$$

For a.e. $\tau \in [0,T]$, we can choose $p(\tau) \in H^1(\Omega)$ as a test function to obtain with (2.15), Definition 2.1.1 and the weak coercivity of $\mathcal{F}_t$ the following estimate

$$\|p(t)\|_2^2 = \|p_0\|_2^2 - 2\int_0^t \mathcal{F}_\tau\big(p(\tau), p(\tau)\big)\, d\tau \le \|p_0\|_2^2 + 2\int_0^t \gamma \|p(\tau)\|_2^2\, d\tau.$$

Since $t \mapsto \|p(t)\|_2$ is continuous, we have by Grönwall's lemma

$$\|p(t)\|_2^2 \leqslant e^{2\gamma t}\, \|p_0\|_2^2, \quad t \in [0,T]. \tag{2.16}$$

In order to verify the $L^2(0,T;H^1(\Omega))$–bound, we notice with Lemma 2.1.2, (2.15) and (2.16) that

$$\begin{aligned}\int_0^T \beta \|p(t)\|_{H^1(\Omega)}^2\, dt &\le \int_0^T \Big(\gamma\|p(t)\|_2^2 - \langle \dot{p}(t), p(t)\rangle_{H'}\Big)\, dt \\ &\le \int_0^T \gamma e^{2\gamma t}\, \|p_0\|_2^2\, dt - \frac{1}{2}\big(\|p(T)\|_2^2 - \|p_0\|_2^2\big) \\ &\le \frac{1}{2} e^{2\gamma T}\, \|p_0\|_2^2.\end{aligned} \tag{2.17}$$

For a $H^1(\Omega)'$–bound, we only use that $\langle \dot{p}(t), \psi\rangle_{H'} = -\mathcal{F}_t(p(t),\psi)$ holds f.a.e. $t \in [0,T]$, and the boundedness of $\mathcal{F}_t$ to obtain

$$\|\dot{p}(t)\|_{H^1(\Omega)'} = \sup_{\psi \in H^1(\Omega)} \frac{|\langle \dot{p}(t), \psi\rangle_{H'}|}{\|\psi\|_{H^1(\Omega)}} \le C\|p(t)\|_{H^1}, \quad t \in [0,T]\backslash N. \tag{2.18}$$

Consequently, the $L^2(0,T;H^1(\Omega)')$–bound follows from the $L^2(0,T;H^1(\Omega))$–bound of p and the proof of b) is complete.
In order to verify uniqueness, assume that $p, \tilde{p} \in C([0,T];L^2(\Omega))$ are both weak solutions to the same initial state p_0. Once again with (2.15) and Lemma 2.1.2 we have

$$\|p(t) - \tilde{p}(t)\|_2^2 = -2\int_0^t \mathcal{F}_\tau\big(p(\tau) - \tilde{p}(\tau), p(\tau) - \tilde{p}(\tau)\big)\, d\tau \le C\int_0^t \|p(\tau) - \tilde{p}(\tau)\|_2^2\, d\tau.$$

Thus, applying Grönwall's lemma gives the assertion and the proof is complete. □

We remark that Theorem 2.1.3 remains valid if we consider controls from merely $L^2(0,T;L^\infty(\Omega))^m$ instead of $L^\infty(\Omega_T)$ by a density argument, and we refer to [5, Theorem 2.2] for a proof. In this case,

$$\|p\|_{W(0,T)} + \|p\|_{L^\infty(0,T;L^2(\Omega))} \le C_F\, C(\|u\|_{L^2(0,T;L^\infty(\Omega))}). \tag{2.19}$$

Next, we show that weak solutions of the Fokker–Planck problem satisfy the typical properties of a PDF. This can be seen as the motivation for considering the flux–zero boundary condition. However, we remark that these boundary conditions can be rigorously derived from the SDE (2.1) in the manner of Section 1.1.

Corollary 2.1.4. *Recall that, according to (F3), p_0 is a probability distribution function, i.e.,*

i) $\int_\Omega p_0(x)\,dx = 1$ *and*

ii) $p_0 \geq 0$ *a.e. on* Ω.

Let p be the unique weak solution in $C([0,T];L^2(\Omega))$. Then, $p(t) \in L^2(\Omega)$ does also have these properties for all $t \in [0,T]$. We say that the Fokker–Planck problem with flux–zero boundary conditions is conservative.

Proof. Due to the flux zero boundary condition, the test function appears only as a gradient in the bilinear form $\mathcal{F}$. Therefore, the conservation of the total probability follows from the definition of a weak solution if we choose $\psi = 1 \in H^1(\Omega)$ as a test function

$$0 = -\int_s^t \mathcal{F}_\tau(p(\tau),\psi)\,d\tau = \int_s^t \langle \dot{p}(\tau),\psi\rangle_{H'}\,d\tau = \int_\Omega p(t)\,dx - \int_\Omega p(s)\,dx, \quad 0 \leq s,t \leq T. \tag{2.20}$$

However, since $\dot{p}(t)$ is only an $H^1(\Omega)'$–function, and since the following argument appears multiple times in this thesis, we carefully prove the last equal sign. First, recall the continuous embedding

$$C^1([0,T];H^1(\Omega)) \subset W^{1,2}(0,T;H^1(\Omega)') \cap L^2(0,T;H^1(\Omega)),$$

and the fundamental theorem of calculus for Banach space valued functions

$$\varphi(t) - \varphi(s) = \int_s^t \dot{\varphi}(\tau)\,d\tau \tag{2.21}$$

a.e.on Ω for all $0 \leq s,t \leq T$, $\varphi \in C^1([0,T];H^1(\Omega))$. Hence, the last equal sign in (2.20) can be proven with the following density argument. Let $\left(p_k\right)_{k\in\mathbb{N}} \subset C^1([0,T];H^1(\Omega))$ with $p_k \to p$ in $W(0,T)$. Now by (2.11), Fubini and (2.21) we have

$$\int_s^t \langle \dot{p}_k(\tau),1\rangle_{H'}\,d\tau = \int_s^t \langle \dot{p}_k(\tau),1\rangle_{L^2(\Omega)}\,d\tau = \int_\Omega \int_s^t \dot{p}_k(\tau)\,d\tau\,dx = \int_\Omega p_k(t)\,dx - \int_\Omega p_k(s)\,dx.$$

Taking the limit on both sides proves the conservation of the total probability.
In order to show the non–negativity of p, we consider its negative part

$$p_- := \min\{p,0\} \in L^2(0,T;H^1(0,T)) \cap L^\infty(0,T;L^2(\Omega)).$$

Note that in general p_- does not belong to $H^1(0,T;H^1(\Omega)')$, nevertheless, an integration–by–parts formula still holds, and we refer to [67] for a proof. This implies f.a.e. $t \in]0,T[$

$$\langle \dot{p_-}(t),p_-(t)\rangle_{H'} = \langle \dot{p}(t),p_-(t)\rangle_{H'}, \text{ and } \mathcal{F}_t(p(t),p_-(t)) = \mathcal{F}_t(p_-(t),p_-(t)).$$

This yields with $p_-(0) = 0$ and the weak coercivity of $\mathcal{F}$ that for every $t \in [0,T]$

$$\frac{1}{2}\|p_-(t)\|_2^2 = \int_0^t \langle \dot{p_-}(\tau),p_-(\tau)\rangle_{H'}\,d\tau = -\int_0^t \mathcal{F}_\tau(p_-(\tau),p_-(\tau))\,d\tau \leq \gamma\int_0^t \|p_-(\tau)\|_2^2\,d\tau.$$

Now, Grönwall's inequality implies that $\|p_-(t)\|_2^2 \leq 0$ which in turn provides $p(t) \geq 0$ a.e. on Ω. □

Let us establish some standard regularity properties for $W(0,T)$–functions. For this purpose, let us recall the following continuous Sobolev embeddings, cf. [1],

$$H^1(\Omega) \hookrightarrow \begin{cases} C^{1/2}(\Omega), & & \text{if } d = 1, \\ L^\eta(\Omega), & \eta \in [1,\infty[& \text{if } d = 2, \\ L^q(\Omega), & q \in [1,\frac{2d}{d-2}[& \text{if } d \geqslant 3. \end{cases} \tag{2.22}$$

Corollary 2.1.5. *(Further regularity of $W(0,T)$–functions)*
Let $d \in \mathbb{N}$ be the dimension of Ω. Then any function in $W(0,T)$ is also in $L^{4/d+2}(\Omega_T)$ and the embedding

$$W(0,T) \Subset L^{\eta}(\Omega_T), \quad 1 \leq \eta < \frac{4}{d} + 2 \tag{2.23}$$

is compact.

Proof. Due to the Gagliardo–Nirenberg interpolation inequality, we obtain the continuous embedding $W(0,T) \hookrightarrow L^{4/d+2}(\Omega_T)$. Next, let $q = \frac{2d}{d-2}$ if $d \geq 3$ and $q = \infty$ else. Let $1 \leq p < q$. Since the Rellich–Kondrachov embedding $H^1(\Omega) \Subset L^p(\Omega)$ is compact, we may apply Aubin–Lions Lemma on

$$H^1(\Omega) \Subset L^p(\Omega) \subset H^1(\Omega)'$$

to obtain the compact embedding $W(0,T) \Subset L^2(0,T;L^p(\Omega)$. Consequently, for any bounded sequence $(z_k) \subset W(0,T)$, we have for a subsequence

$$z_k \to z \quad \text{in } L^2(0,T;L^p(\Omega)) \quad \text{and} \quad |z_k - z| \text{ is uniformly bounded in } L^\infty(0,T;L^2(\Omega)).$$

With a standard interpolation estimate for Bochner spaces, we obtain

$$\|z_k - z\|_{\tau,r} \leq \|z_k - z\|_{\infty,2}^{1-\alpha} \|z_k - z\|_{2,p}^{\alpha} \to 0,$$

where $\frac{1}{\tau} = \frac{1-\alpha}{\infty} + \frac{\alpha}{2}$ and $\frac{1}{r} = \frac{1-\alpha}{2} + \frac{\alpha}{p}$. Rearranging both equations to τ and r with $\tau = r$ yields the assertion as p tends to q. □

When we analyze FP optimal control problems, we have to consider the Fréchet derivatives of the control–to–state map. This operator will be given implicitly by an inhomogeneous Fokker–Planck problem, and therefore, we have to investigate existence and regularity in the following section.

2.2 The inhomogeneous Problem – obtaining uniform bounds with a De Giorgi iteration

> *When you change the way you look at things, the things you look at change.*
>
> Max Planck, 1858 – 1947

In preparation of our analysis of optimality conditions, we discuss an inhomogeneous FP equation with a right–hand side belonging to the space $L^2(0,T;H^1(\Omega)')$. The main result of this section is the L^∞–estimate given in Theorem 2.2.3 below, which is essential for the upcoming analysis of the FP ensemble optimal control problem in the case of time–space dependent controls. Furthermore, we present an L^∞–estimate for an inhomogeneous parabolic problem with right–hand side belonging to $L^\infty(\Omega_T)$, which is needed for the adjoint problem.

Corollary 2.2.1. *Let $g \in L^2(0,T;H^1(\Omega)')$, $z_0 \in L^2(\Omega)$ and $u \in L^2(0,T;L^\infty(\Omega))^m$. Then there exists a unique weak solution $z \in W(0,T)$ of the inhomogeneous Fokker–Planck problem in the sense that there exists a null set $N \subset [0,T]$ with*

$$\langle \dot{z}(t), \psi \rangle_{H'} + \mathcal{F}_t(z(t), \psi) = \langle g(t), \psi \rangle_{H'}, \quad t \in [0,T]\backslash N,\ \psi \in H^1(\Omega),$$

with initial condition $z(0) = z_0$ a.e. on Ω. Additionally, there exists a constant $C = C_{\mathrm{F}} C_u$, where C_u depends continuously only on $\|u\|_{L^2 L^\infty}$, such that

$$\|z\|_{L^\infty(0,T;L^2(\Omega))} + \|z\|_{L^2(0,T;H^1(\Omega))} + \|\dot{z}\|_{L^2(0,T;H^1(\Omega)')} \leq C\big(\|z_0\|_{L^2}^2 + \|g\|_{L^2(0,T;H^1(\Omega)')}\big).$$

Proof. Due to the linearity of the Fokker–Planck equation, the proof can be easily deduced from the proof of Theorem 2.1.3. □

The following L^∞–estimate is crucial for the second–order analysis and is shown with a De Giorgi iteration. For the convenience of the reader, we state the so–called De Giorgi lemma; a proof can be found in [68, Lemma 4.1.1].

Lemma 2.2.2. *(De–Giorgi Iteration)*
Let $\lambda_0 \geq 0$. Let $\varphi : [\lambda_0, \infty[\to [0, \infty[$ be a non–increasing function, satisfying for some constants $M, \alpha > 0$, $\beta > 1$ the estimate

$$\varphi(m) \leq \left(\frac{M}{m-\lambda}\right)^\alpha \varphi(\lambda)^\beta$$

for all $m > \lambda \geq \lambda_0$. Then, there exists $C > 0$ such that for all $\lambda \geq \lambda_0 + C$

$$\varphi(\lambda) = 0.$$

Although new, the following result is known to be true for similar parabolic equations, and we were able to use the available techniques of the proof to our case; see [14] and [68, Theorem 4.2.2]. We remark that we impose $u \in L^\infty(\Omega_T)$ and to the best of our knowledge, merely $u \in L^2(0,T;L^\infty(\Omega))$ is not sufficient for an L^∞–estimate.

Theorem 2.2.3. *(L^∞–estimates for the inhomogeneous Fokker–Planck problem)*
Let $z_0 \in L^\infty(\Omega)$, $u \in L^\infty(\Omega_T)$ and let $z \in W(0,T) \cap C([0,T];L^2(\Omega))$ be the unique weak solution of the inhomogeneous Fokker–Planck problem

$$\langle \dot{z}, \cdot \rangle_{H'} + \mathcal{F}(z, \cdot) = \langle \mathcal{G}, \cdot \rangle_{H'}, \quad \textit{in } L^2(0,T;H^1(\Omega)')$$

with $z(0) = z_0$ a.e. on Ω. Let the source term be of the form

$$\langle \mathcal{G}_t, \psi \rangle_{H'} := \int_\Omega \Big(g_1(t,x)\, \psi(x) + g_2(t,x) \cdot \nabla\psi(x) \Big)\, dx, \quad t \in [0,T],\ \psi \in H^1(\Omega), \tag{2.24}$$

where $g_1 \in L^q(\Omega_T)$ and $g_2 \in L^q(\Omega_T)^d$ with $q > d+2$. Furthermore, let $z \in L^q(\Omega_T)$. Then, $z \in L^\infty(\Omega_T)$ and there exist some constant $C = C_F C_u > 0$, where C_u depends continuously only on $\|u\|_\infty$, such that

$$\|z(t)\|_\infty \leq e^{Ct}\|z_0\|_\infty + C\left(\|g_1\|_q + \|g_2\|_q + \|z\|_q\right), \qquad t \in [0,T]. \tag{2.25}$$

We remark that if $d \in \{1,2\}$, then $z \in L^q(\Omega_T)$ due to Corollary 2.1.5.

Proof. For any $\gamma > 0$, $\lambda > \|z_0\|_\infty$, we define the $C([0,T];L^2(\Omega))$–functions

$$f(t,x) := e^{-\gamma t} z(t,x), \qquad f_\lambda(t,x) := \max\{f(t,x) - \lambda, 0\}, \quad (t,x) \in [0,T] \times \Omega.$$

Notice that $f \in W(0,T)$, hence, f_λ is non–negative on Ω_T, positive on the measurable set

$$M_\lambda := \{(t,x) \in \Omega_T \ : \ f(t,x) > \lambda\}$$

and an integration–by–parts formula holds, cf. [67]. We remark that the $(d+1)$–dimensional volume of M_λ does not depend on the choice of the pointwise defined representative of z. Furthermore, we can assume that $\operatorname{vol} M_\lambda > 0$ for all $\lambda > \|z_0\|_\infty$, otherwise the assertion is already shown.
Step 1: For a.e. $t \in [0,T]$, we observe that

$$\frac{1}{2}\frac{d}{dt}\left(\|f_\lambda(t)\|_2^2\right) = \langle \dot{f}(t), f_\lambda(t) \rangle = -\gamma \int_\Omega f(t) f_\lambda(t)\, dx - \mathcal{F}_t(f(t), f_\lambda(t)) + \mathcal{G}_t(f_\lambda(t)), \tag{2.26}$$

since p solves $\langle p, \cdot\rangle = -\mathcal{F}_t(p,\cdot)$ in $L^2(0,T;H^1(\Omega)')$ and $f_\lambda(t) \in H^1(\Omega)$. Due to (F2), we find that a.e. on $[0,T]$ it holds that

$$\begin{aligned} -\mathcal{F}(f,f_\lambda) &= -\int_\Omega \left(\sum_{i,j=1}^d a_{ij}\partial_{x_i} f_\lambda \partial_{x_j} f_\lambda - fb\cdot\nabla f_\lambda\right) dx \\ &\le -\theta\int_\Omega |\nabla f_\lambda|^2\,dx + \int_\Omega fb\cdot\nabla f_\lambda\,dx. \end{aligned} \tag{2.27}$$

In the first step, we have used the fact that $f_\lambda = 0$ on $\Omega_T \backslash M_\lambda$ and $\nabla f_\lambda = \nabla f$ on M_λ.
Now, since $\lambda > \|z_0\|_\infty$, we have $\|f_\lambda(0)\|_2 = 0$. Combining (2.26) and (2.27), and integrating with respect to t yields

$$\begin{aligned} \frac{1}{2}\|f_\lambda(t)\|_2^2 &= -\gamma\int_0^t\int_\Omega f(s,x)f_\lambda(s,x)\,ds\,dx - \int_0^t \mathcal{F}_s(f(s),f_\lambda(s))\,ds + \int_0^t \mathcal{G}_s(f_\lambda(s))\,ds \\ &\le \int_0^t \left(-\gamma\lambda\int_\Omega f_\lambda\,dx - \gamma\|f_\lambda\|_2^2 - \theta\|\nabla f_\lambda\|_2^2\right) ds \\ &\quad + \int_0^t\int_\Omega \left(g_1 f_\lambda + (g_2+fb)\cdot\nabla f_\lambda\right)\,dx\,ds, \end{aligned} \tag{2.28}$$

where we suppress the arguments of the functions in the last step for the sake of clarity. We use the ε-Young inequality to obtain on M_λ

$$(g_2+fb)\cdot\nabla f_\lambda \le \frac{4}{\varepsilon}\left(|fb|^2+|g_2|^2\right) + 2\varepsilon|\nabla f_\lambda|^2, \qquad g_1 f_\lambda \le \frac{4}{\varepsilon}g_1^2 + \varepsilon f_\lambda^2.$$

Since $-\gamma\lambda\int_\Omega f_\lambda(t,x)\,dx$ is non–positive, we obtain with (2.28) the following inequality

$$\begin{aligned} \frac{1}{2}\|f_\lambda(t)\|_2^2 \le& \int_0^t \left((\varepsilon-\gamma)\|f_\lambda(s)\|_2^2 + (2\varepsilon-\theta)\|\nabla f_\lambda(s)\|_2^2\right) ds \\ &+ \frac{4}{\varepsilon}\left(\|g_1\|_{L^2(M_\lambda)}^2 + \|g_2\|_{L^2(M_\lambda)}^2 + \|b\|_\infty^2\|f\|_{L^2(M_\lambda)}^2\right). \end{aligned}$$

Next, the choice $\varepsilon = \theta/4$, $\gamma = \theta/2$ results in both $(2\varepsilon - \theta)$ and $(\varepsilon-\gamma)$ being negative; thus we arrive at

$$\|f_\lambda\|_{\infty,2}^2 + \|f_\lambda\|_{2,H^1}^2 \le C\left(\|g_1\|_{L^2(M_\lambda)}^2 + \|g_2\|_{L^2(M_\lambda)}^2 + \|f\|_{L^2(M_\lambda)}^2\right). \tag{2.29}$$

Step 2: Since $f_\lambda \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega))$, Corollary 2.1.5 yields $f_\lambda \in L^\eta(\Omega_T)$ for all

$$1 \le \eta < 2 + \frac{4}{d},$$

and we can bound

$$\|f_\lambda\|_\eta \le C\|f_\lambda\|_{\infty,2}^{1-\alpha}\|f_\lambda\|_{2,H^1}^\alpha \le C\left(\|f_\lambda\|_{\infty,2}^2 + \|f_\lambda\|_{2,H^1}^2\right)^{1/2},$$

where we used Young's inequality for the second estimate. Next, we apply Hölder's inequality with the indicator function and exponent $\kappa' := q/2$ with dual κ to obtain for $i = 1,2$

$$\begin{aligned} \int_{M_\lambda} |g_i|^2\,dt\,dx &\le (\operatorname{vol} M_\lambda)^{1/\kappa}\left(\int_{M_\lambda} |g_i|^{2\kappa'}\,dt\,dx\right)^{1/\kappa'}, \\ \int_{M_\lambda} |f|^2\,dt\,dx &\le (\operatorname{vol} M_\lambda)^{1/\kappa}\left(\int_{M_\lambda} |f|^{2\kappa'}\,dt\,dx\right)^{1/\kappa'}. \end{aligned}$$

This implies with (2.29) and $q = 2\kappa'$ the estimate

$$\|f_\lambda\|_\eta \le C\,(\operatorname{vol} M_\lambda)^{1/(2\kappa)}\,G, \qquad \text{with } G := \left(\|g_1\|_{L^q(\Omega_T)} + \|g_2\|_{L^q(\Omega_T)} + \|f\|_{L^q(\Omega_T)}\right). \tag{2.30}$$

Step 3: In this step, we bring the previous results together and consider the well–defined, non–increasing function $\varphi(\lambda) := \operatorname{vol} M_\lambda$, defined for $\lambda \in [\,\|f_0\|_\infty, \infty[\,$. Now let $m > \lambda > \|f_0\|_\infty$. Hence, $M_m \subset M_\lambda$ and on M_m, it holds that $m - \lambda \le f - \lambda = f_\lambda$. Thus, by (2.30), we obtain

$$\varphi(m)(m-\lambda)^\eta = \int_{M_m} (m-\lambda)^\eta \le \int_{M_\lambda} f_\lambda^\eta \, dt\, dx \le C\,(\varphi(\lambda))^{\eta/(2\kappa)}\, G^\eta,$$

and therefore, it holds that $\varphi(m) \le C\,(G/(m-\lambda))^\eta\,(\varphi(\lambda))^{\eta/(2\kappa)}$.
In order to apply the De–Giorgi Iteration, we must verify that the exponent $\eta/(2\kappa)$, which is β in Lemma 2.2.2, is greater than 1.
We recall that $\kappa = \frac{\kappa'}{\kappa'-1} = \frac{q}{q-2}$ and $1 \le \eta < 2 + \frac{4}{d}$, and consequently, we obtain the condition $\frac{(2+4/d)(q-2)}{2q} > 1$. Rearranging for q, we obtain the condition $q > d+2$. Thus, we may apply Lemma 2.2.2, which implies $\operatorname{vol} M_m = 0$ for $m \ge \|f_0\|_\infty + CG$.
Analogously, we can show that the set $M_\lambda^- := \{(t,x) \in \Omega_T \,:\, f(t,x) < -\lambda\}$ has measure zero for sufficiently large λ by considering $f_\lambda^- := (f+\lambda)_-$ instead of f_λ, which yields the desired lower bound of f. Combining both results, we have shown that

$$\|f\|_\infty \le \|f_0\|_\infty + C(\|g_1\|_q + \|g_2\|_q + \|f\|_q).$$

Since $f(t,x) := e^{-\gamma t} z(t,x)$ we have proven estimate (2.25) under the assumption that $z \in L^q(\Omega_T)$. Now, for the case $d \in \{1,2\}$, Corollary 2.1.5 states that $W(0,T) \subset L^q(\Omega_T)$. This continuous embedding and the fact that $z \in W(0,T)$ concludes the proof. □

We remark that the assumption on the exponent q can be weakened if $g_2 = 0$. For the analysis of optimality conditions for the FP control problem, we need L^∞–bounds for the adjoint problem. The existence of solutions in $W(0,T)$ is established in Section 3.4.

Theorem 2.2.4. *(L^∞–estimates for the adjoint problem)*
Let $y_0 \in L^\infty(\Omega)$, $g, u \in L^\infty(\Omega_T)$ and let $y \in W(0,T) \cap C([0,T]; L^2(\Omega))$ be the unique weak solution of the following problem

$$\langle \dot y, \cdot\rangle_{H'} + \mathcal{F}(\cdot, y) = \langle g, \cdot\rangle_{H'}, \quad \textit{in } L^2(0,T; H^1(\Omega)')$$

with $y(0) = y_0$ a.e. on Ω. Let $q > d+2$. If $g \in L^q(\Omega_T)$ and $\nabla y \in L^q(\Omega_T)^d$, then $y \in L^\infty(\Omega_T)$ and there exists some $C = C_F C_u > 0$ such that

$$\|y(t)\|_\infty \le e^t \|y_0\|_\infty + C(\|g\|_\infty + \|\nabla y\|_q), \quad t \in [0,T]. \tag{2.31}$$

Proof. The proof can be done analogously to the one of Theorem 2.2.3; notice that the only change is y appearing in the second argument of $\mathcal{F}(\cdot,\cdot)$ instead in the first one as in the FP problem. Therefore, we similarly define $f(t,x) := e^{-\gamma t} y(t,x)$ and $f_\lambda(t,x) := (f(t,x) - \lambda)_+$, and we see that equation (2.26)–(2.27) changes to

$$\frac{1}{2}\frac{d}{dt}\left(\|f_\lambda(t)\|_2^2\right) = \langle \dot f(t), f_\lambda(t)\rangle = \int_\Omega \big(g(t) f_\lambda(t) - \gamma f(t) f_\lambda(t)\big)\, dx - \mathcal{F}_t(f_\lambda(t), f(t))$$

f.a.e. $t \in [0,T]$ and

$$-\mathcal{F}(f_\lambda, f) \le -\theta \int_\Omega |\nabla f_\lambda|^2\, dx + \int_\Omega f_\lambda b \cdot \nabla f\, dx, \quad \text{a.e. on } [0,T].$$

Consequently, estimate (2.28) becomes f.a.e. $t \in [0,T]$

$$\begin{aligned}\frac{1}{2}\|f_\lambda(t)\|_2^2 \le \int_0^t \left(-\gamma\lambda\int_\Omega f_\lambda(s,x)\,dx - \gamma\|f_\lambda(s)\|_2^2 - \theta\|\nabla f_\lambda(s)\|_2^2\right)\,ds \\ + \int_0^t\int_\Omega f_\lambda(s,x)(b(s,x)\cdot\nabla f(s,x) + g(s,x))\,dx\,ds.\end{aligned}$$

We apply the ε–Young inequality to estimate $f_\lambda(b\cdot\nabla f + g) \le \varepsilon f_\lambda^2 + \frac{4}{\varepsilon}(b\cdot\nabla f + g)^2$ on M_λ, and since $\gamma\lambda\int_\Omega f_\lambda\,dx \ge 0$, we obtain

$$\begin{aligned}\frac{1}{2}\|f_\lambda(t)\|_2^2 &\le \int_0^t \left(-\gamma\|f_\lambda(s)\|_2^2 - \theta\|\nabla f_\lambda(s)\|_2^2\right)\,ds \\ &\quad + \varepsilon\int_0^t\int_\Omega |f_\lambda(s,x)|^2\,ds\,dx + \frac{4}{\varepsilon}\int_0^t\int_\Omega (b(s,x)\cdot\nabla f(s,x) + g(s,x))^2\,dx\,ds \\ &\le \int_0^t \left((\varepsilon-\gamma)\|f_\lambda(s)\|_2^2 - \theta\|\nabla f_\lambda(s)\|_2^2\right)\,ds + \frac{8}{\varepsilon}\left(\|b\|_\infty^2\|\nabla f\|_2^2 + \|g\|_2^2\right).\end{aligned}$$

We recall that we use the same notation $\|\cdot\|_p$ for the L^p–norm over Ω and Ω_T. Once again, we choose ε and γ such that $(\varepsilon-\gamma)$ is negative – in contrast to the proof of Theorem 2.2.3, we may simply choose $\gamma = 1$ and $\varepsilon = 1/2$ – and we arrive at

$$\|f_\lambda\|_{\infty,2}^2 + \|f_\lambda\|_{2,H^1}^2 \le C\left(\|g\|_{L^2(M_\lambda)}^2 + \|\nabla f\|_{L^2(M_\lambda)}^2\right).$$

Next, we follow step 2 of the proof of Theorem 2.2.3 and obtain (2.30) with $G := \|g\|_q + \|\nabla f\|_q$. Step 3 can be done completely analogously, and we arrive at

$$\|f\|_\infty \le \|f_0\|_\infty + C(\|g\|_q + \|\nabla f\|_q).$$

The assertion follows from the fact that $f(t,x) = e^{-t}y(t,x)$, which implies $\|f_0\|_\infty = \|y_0\|_\infty$. This concludes the proof. □

Let us remark that the estimate (2.31) is not optimal, since the choice $\gamma = 1$ and $\varepsilon = 1/2$ in the proof have not been optimal.

2.3 Higher regularity of solutions to parabolic problems

In this section, we establish higher regularity of weak solutions to the Fokker–Planck problem and related parabolic problems under all assumptions (F1)–(F7). Let us state a well–known result from [34, Theorem 5] for a parabolic problem with Dirichlet boundary conditions. Let

$$y \in L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$$

be a weak solution to the inhomogeneous problem

$$\partial_t y + Ly = g \quad \text{on } \Omega_T$$

with r.h.s. g, initial condition $y(0) = y_0$ on Ω and Dirichlet boundary condition $y(t,\cdot) = 0$ in the trace sense on $\partial\Omega$ for a.e. $t \in [0,T]$. Then, if the coefficients of L are sufficiently smooth, $y_0 \in H_0^1(\Omega)$ and $g \in L^2(\Omega_T)$, the weak solution y enjoys the higher regularity

$$y \in L^2(0,T;H^2(\Omega)) \cap C([0,T];H_0^1(\Omega)), \quad \partial_t y \in L^2(\Omega_T).$$

To the best of our knowledge, an analogous result does not exist in general for our FP problem with the flux–zero boundary conditions (2.4). In [59, Section 3], it is claimed that higher regularity can be shown by a classical bootstrap argument, but no proof is given, and it seems difficult to verify the claim in this general setting. In the works of [5,13], the authors deduce higher regularity by a different approach, which is discussed next. In both papers, the idea is that the Fokker–Planck problem with flux zero boundary conditions can be rewritten as a heat equation with Neumann–boundary conditions under the condition (F4) on M, c and u. This is demonstrated in the following lemma.

Lemma 2.3.1. *Let the conditions (F1)–(F7) hold. Let p be a weak solution of the FP problem.*

a) If (F6 i) holds, then p is a weak solution to the following linear heat problem

$$\begin{aligned} \partial_t p - a\Delta p &= f_1 && \text{on } \Omega_T, \\ p(0) &= p_0 && \text{on } \Omega, \\ \nabla p \cdot \hat{n} &= 0 && \text{on } [0,T] \times \partial\Omega, \end{aligned}$$

where the r.h.s. f_1 is defined as

$$f_1(t,x) := \operatorname{div}\big(p(t,x)B[u](t,x)\big), \quad (t,x) \in \Omega_T.$$

b) If (F6 ii) holds, i.e., c has a potential V, then $w := e^{V/a}p$ is a weak solution of

$$\begin{aligned} \partial_t w - a\Delta w &= f_2 && \text{on } \Omega_T, \\ w(0) &= p_0 && \text{on } \Omega, \\ \nabla w \cdot \hat{n} &= 0 && \text{on } [0,T] \times \partial\Omega, \end{aligned}$$

where we have possibly changed V up to a constant such that $V(0,\cdot) = 0$ a.e. on Ω. The r.h.s. f_2 is defined for $(t,x) \in \Omega_T$ as

$$f_2(t,x) := -c(t,x) \cdot \nabla w(t,x) - \frac{1}{a} w(t,x)\, c(t,x)^\top M(t,x)u(t,x) + \operatorname{div}\big(w(t,x)M(t,x)u(t,x)\big).$$

Proof. Let (F6 i) be fulfilled, that is, f.a.e. $t \in [0,T]$ we have $c(t,\cdot) \cdot \hat{n} = 0$ a.e. on $\partial\Omega$. Due to this and (F5), we have $(pB[u]) \cdot n = 0$ on $[0,T] \times \partial\Omega$. Hence, for any test function $\psi \in H^1(\Omega)$, it holds that

$$\int_{\Omega_T} pB[u] \cdot \nabla\psi \, dx = -\int_{\Omega} \operatorname{div}\big(pB[u]\big)\psi \, dx = -\langle f_1, \psi \rangle_{L^2(\Omega)}.$$

This proves the first claim.
Next, let (F6 ii) hold and define $w := e^{V/a}p$; notice that p and w enjoy the same regularity, up to the regularity of V. Obviously, $w(0) = e^{V(0)/a}p(0) = p_0$ a.e. on Ω. With the chain rule, we compute the weak derivatives

$$\nabla p = e^{-V/a}\nabla w - e^{-V/a}\nabla V \, \frac{w}{a} = e^{-V/a}\Big(\nabla w - \frac{c\,w}{a}\Big), \quad \text{a.e. on } \Omega_T.$$

We insert the formulas for ∇p and p into the flux–zero boundary conditions to obtain

$$0 = \big(\nabla p + pB[u]\big) \cdot \hat{n} = a\, e^{-V/a}\nabla w \cdot \hat{n}, \quad \text{a.e. on } [0,T] \times \partial\Omega.$$

Since a and $e^{-V/a}$ are positive, this implies

$$\nabla w \cdot \hat{n} = 0 \quad \text{a.e. on } [0,T] \times \partial\Omega.$$

Lastly, the formula for f_2 is obtained by computing $\operatorname{div}(e^{V/a}(\nabla p + cp/a))$ and applying the boundary conditions,

$$a \int_\Omega \nabla w \cdot \nabla \psi \, dx = a \int_\Omega e^{V/a} \Big(\nabla p + \frac{cp}{a} \Big) \cdot \nabla \psi \, dx = -\langle f_2, \psi \rangle_{L^2(\Omega)}.$$

This shows the second claim. □

Higher regularity of p depends on the regularity of f_1 and f_2, which in turn depends on the set of admissible controls under consideration. In general, if u is space dependent, derivatives of u appear in f_1 and f_2. Therefore, one would need a set of admissible controls such that the spatial derivatives of u are essentially bounded, in order for f_1 and f_2 to be in $L^2(\Omega_T)$. Therefore, the following Lemma is the key for higher regularity but in our setting only applicable in the case of only time–dependent controls.

Lemma 2.3.2. *Let the conditions (F1)–(F7) hold. Let the control $u \in L^\infty(0,T)^m$ be only time–dependent.*

a) The functions f_1 and f_2 defined in Lemma 2.3.1 are in $L^2(\Omega_T)$, and p satisfies (depending on the case (F6 i) or (F6 ii) for $j = 1, 2$)

$$\|p\|_{W(0,T)} + \|p\|_{C([0,T];H^1(\Omega))} \leq C_{\mathrm{F}*} \big(\|f_j\|_{L^2(\Omega_T)} + \|p_0\|_{H^1(\Omega)} \big).$$

b) The mapping $f \mapsto p$ is also continuous from $L^2(\Omega_T)$ to $C^{1/2}([0,T];L^2(\Omega)) \cap C([0,T];H^1(\Omega))$ and compact from $L^2(\Omega_T)$ to $C([0,T];L^2(\Omega))$, where p is a weak solution to

$$\begin{aligned} \partial_t p - a\Delta p &= f && \text{on } \Omega_T, \\ p(0) &= p_0 && \text{on } \Omega, \\ \nabla p \cdot \hat{n} &= 0 && \text{on } [0,T] \times \partial\Omega. \end{aligned}$$

Proof. The proof is given in the Appendix of [5]. □

Let us summarize the issue of higher regularity of solutions to the FP problem. When we consider second–order conditions of an optimization problem constrained by the FP problem, it turns out that solutions merely in $W(0,T)$ are not enough to prove certain statements.

In the case of space–dependent controls, we will see quickly that it is convenient to work with essentially bounded solutions. For that purpose, we have established the L^∞–estimate in the previous section; however, this restricts us to dimensions $d \in \{1,2,3\}$ of Ω.

In the case of only time–dependent controls, we have a second option and can avoid the necessity of essentially bounded solutions. This is done by rewriting the FP problem as a linear heat equation with a r.h.s. and then apply Lemma 2.3.2. This gives weak solutions in $C([0,T];H^1(\Omega))$, a very useful compactness result and holds in all dimensions $d \in \mathbb{N}$ of Ω. When we are interested in the numerical analysis for the FP problem, we need to obtain even higher regularity than $C([0,T];H^1(\Omega))$. For this purpose, let us introduce Bessel potential spaces and Sobolev–Slobodeckij spaces, which allow a finer classification of functions from Sobolev spaces.

Definition 2.3.3. *(Bessel potential space and Sobolev–Slobodeckij space)*
Let $n \in \mathbb{N}$ and let $D \subset \mathbb{R}^n$ be a bounded domain.

a) For $f \in L^2(\mathbb{R}^n)$, we denote by

$$\hat{f}(\xi) := \int_{\mathbb{R}^n} f(x) e^{-2\pi i x \cdot \xi} \, dx, \quad \xi \in \mathbb{R}^n$$

the Fourier transform of f. The Fourier transform $\hat{\cdot}$ is an isomorphism on $L^2(\mathbb{R}^n)$, and its inverse is denoted by $\check{f}$.

b) For $q \in [1,\infty[$ and $s \in \mathbb{R}$, we define the Bessel potential space

$$H_q^s(\mathbb{R}^n) := \Big\{ f \in L^q(\mathbb{R}^n) : \check{g}_s \in L^q(\mathbb{R}^n) \quad \textit{with} \quad g_s(\xi) := (1+|\xi|^2)^{s/2}\hat{f}(\xi), \quad \xi \in \mathbb{R}^n \Big\}$$

with the norm

$$\|f\|_{H_q^s(\mathbb{R}^n)} := \|\check{g}_s\|_{L^q(\mathbb{R}^n)}.$$

Furthermore, we define

$$H_q^s(D) := \big\{ f_{|D} : f \in H_q^s(\mathbb{R}^n) \big\}$$

with the norm

$$\|f\|_{H_q^s(D)} := \inf \Big\{ \|h\|_{H_q^s(\mathbb{R}^n)} : h \in H_q^s(\mathbb{R}^n) \quad \textit{and} \quad h_{|D} = f \Big\}.$$

c) For $q \in [1,\infty[$ and $s \in]0,1[$, we define the Sobolev–Slobodeckij space

$$W^{s,q}(D) := \Big\{ f \in L^q(D) : \int_D \int_D \frac{|f(x)-f(y)|^q}{|x-y|^{n+sq}}\,dx\,dy < \infty \Big\}$$

with the norm

$$\|f\|_{W^{s,q}(D)} := \Big(\int_D |u(x)|^q\,dx + \int_D \int_D \frac{|f(x)-f(y)|^q}{|x-y|^{n+sq}}\,dx\,dy \Big)^{1/q}$$

We remark that the Sobolev–Slobodeckij space can be seen as the extension of the Sobolev space $W^{k,q}(D)$, $k \in \mathbb{N}$ for fractional derivatives. In view of that, the Sobolev–Slobodeckij space $W^{k+s,q}(D)$ can be defined as the set of functions from $W^{k,p}(D)$ with weak k–th derivatives in $W^{s,q}(D)$. Moreover, the Bessel potential space can be seen as the extension of the Hilbert space $H^s(D)$, $s \in \mathbb{R}$, for other exponents than $q = 2$, and we notice that

$$W^{s,2}(D) = H_2^s(D) = H^s(D).$$

Furthermore, we collect the following basic properties of the Bessel potential space and Sobolev–Slobodeckij space. We refer the reader to [63, Remark 1.96] and [26, 55] for the well–definedness of the Bessel spaces and Sobolev–Slobodeckij spaces and for the proof of the following lemma.

Lemma 2.3.4. *Let $n \in \mathbb{N}$ and let $D \subset \mathbb{R}^n$ be a bounded domain.*

a) For all $s \in]0,1[$ and $q \in]1,2]$, it holds that

$$W^{s,q}(D) \hookrightarrow H_q^s(D).$$

b) For all $s \in]0,1[$ and $q \in [2,\infty[$, it holds that

$$H_q^s(D) \hookrightarrow W^{s,q}(D).$$

c) If $sq < n$, then for any $q^* \in [1, \frac{nq}{n-sq}]$ it holds that

$$H_q^s(D) \hookrightarrow L^{q^*}(D).$$

We remark that this is the extension of the Gagliardo–Nirenberg–embedding $W^{k,q}(D) \hookrightarrow L^{q^*}(D)$ for $k \in \mathbb{N}$.

d) If $sq > n$ and if $l + \alpha = s - n/q$ with $\alpha \in]0,1[$ and $l \in \mathbb{N}_0$, it holds that

$$H_q^s(D) \hookrightarrow C^{l,\alpha}(D),$$

which is the extension of Morrey's embedding into Hölder spaces $W^{k,q}(D) \hookrightarrow C^{l,\alpha}(D)$.

The Bessel potential spaces play a crucial role for the analysis of the FP optimal control problem, due to the following result on maximal regularity for parabolic problems from [57].

Theorem 2.3.5. *(On maximal L^p–regularity)*
Let $1 < q < \infty$, $q \neq 3/2$, $q \neq 3$ and let $p \in W(0,T)$ be the unique weak solution of the inhomogeneous heat equation with Neumann boundary conditions

$$\partial_t p(t,x) - \Delta p(t,x) = f(t,x) \qquad (t,x) \in [0,T] \times \Omega, \tag{2.32}$$

$$p(0,x) = p_0(x) \qquad x \in \Omega, \tag{2.33}$$

$$\nabla p(t,x) \cdot \hat{n}(x) = 0 \qquad (t,x) \in [0,T] \times \partial\Omega. \tag{2.34}$$

Then, p has the higher regularity

$$p \in H^1_q(0,T;L^q(\Omega)) \cap L^q(0,T;H^2_q(\Omega))$$

if and only if

$$f \in L^q(\Omega_T), \quad p_0 \in W^{2-2/q}_q(\Omega),$$

and the compatibility condition $\nabla p_0 \cdot \hat{n} = 0$ on $\partial\Omega$ holds if $q > 3$.

Furthermore, by standard interpolation arguments, we obtain that

$$H^1_q(0,T;L^q(\Omega)) \cap L^q(0,T;H^2_q(\Omega)) \hookrightarrow H^{1-\theta}_q(0,T;H^{2\theta}_q(\Omega)), \quad \theta \in]0,1[\,.$$

In order to gain higher regularity for FP solutions with Theorem 2.3.5, we need to consider the regularity of the r.h.s. f; for simplicity, let us consider the case a) in Lemma 2.3.2 with $f = \operatorname{div}(p\,B[u])$. Thus, the regularity of f strongly depends on the space of admissible controls under investigation.
Let us consider the simple case of only time–dependent controls from $L^\infty(0,T)^m$ with (F1)–(F7). Since $B[u] = Mu + c$ with

$$M \in L^\infty(0,T;W^{1,\infty}(\Omega))^{d\times m}, \quad c \in L^\infty(0,T;W^{1,\infty}(\Omega))^d$$

we obtain that $\operatorname{div} B[u] \in L^\infty(\Omega_T)$. Thus, $p \in W(0,T)$ implies

$$f = \operatorname{div}(p\,B[u]) \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;L^{p^*}(\Omega)) \hookrightarrow L^\eta(\Omega_T),$$

see Corollary 2.1.5 for the latter embedding and the definition of p^* and η. Consequently, Theorem 2.3.5 lifts the regularity of p from $W(0,T)$ to $H^1_\eta(0,T;L^\eta(\Omega)) \cap L^\eta(0,T;H^2_\eta(\Omega))$. Since p appears in f, we may say that p lifted its regularity on its own, and we can repeat this argument – often referred to as bootstrap argument – as long as we obtain improvement in the regularity of f.

2.4 The control–to–state map, Fréchet differentiability and the linearized state equation

In this section, we introduce and analyze the mapping of a control u to its corresponding state p solving our Fokker–Planck problem. Moreover, we prove compactness of this FP control–to–state map for time– and time–space dependent controls. In the case of only time–dependent controls, we will exploit the higher regularity of p, established in the previous section. For time–space dependent controls, we rely on L^∞–estimates given by Theorem 2.2.3, and therefore, we are restricted to $d \in \{1,2,3\}$ in that case.

Let the assumptions (F1)–(F3) from Chapter 2 hold and let $-\infty < u^{\min} < u^{\max} < \infty$, where $C_{\mathrm{ad}} > 0$ stands for a generic constant that depends continuously only on $|u^{\min}|$ and $|u^{\max}|$. We define the set of admissible controls for the time–space and time–dependent case as follows.

$$\begin{aligned} U_{\mathrm{ad}} &:= \left\{ u \in L^\infty(\Omega_T)^m \,:\, u^{\min} \le u_i \le u^{\max}, \quad \text{a.e. on } \Omega_T,\ i = 1, \dots, m \right\}, \\ U_{\mathrm{ad}}^T &:= \left\{ u \in L^\infty(0,T)^m \,:\, u^{\min} \le u_i \le u^{\max}, \quad \text{a.e. on } [0,T],\ i = 1, \dots, m \right\}. \end{aligned} \tag{2.35}$$

(U1) For time–space dependent controls, we introduce the Hilbert spaces

$$Y_1 := L^2(\Omega_T)^m, \quad Y_2 := L^2(0,T; H_0^1(\Omega))^m, \quad Y_3 := H^1(\Omega_T)^m$$

and the admissible sets

$$U_{\mathrm{ad}}^j := U_{\mathrm{ad}} \cap Y_j, \quad \text{for } j \in \{1, 2, 3\}.$$

(U2) For only time–dependent controls, we analogously define

$$Y_T := L^2(0,T)^m \text{ and } Y_H := H^1(0,T)^m,$$

and admissible sets U_{ad}^T and $U_{\mathrm{ad}}^{T,H} := U_{\mathrm{ad}}^T \cap H_0^1(0,T)^m$.

Throughout this chapter, we use the symbol $\mathcal{U}$ to represent any of these set of admissible controls and the symbol Y for any Hilbert space from above. We remark that the admissible sets are convex, bounded and closed w.r.t. the corresponding norm. Furthermore, we notice that the interior of U_{ad}^j and $U_{\mathrm{ad}}^{T,j}$ with respect to the L^∞–norms are non–empty, which gives meaning to Fréchet differentiability on the admissible sets.

Obviously, one could generalize the constant box–constraints to vector valued functions $u^{\min}, u^{\max} : \overline{\Omega_T} \to \mathbb{R}^m$ or $u^{\min}, u^{\max} : [0,T] \to \mathbb{R}^m$ that are measurable and bounded functions such that the interior of the admissible set is non–empty. We are not pursuing that generalization, since we rather prefer to keep the notations simple.

The control u is from here on added to the notation of the bilinear flux and $\mathcal{F}[u]$ is written instead of just $\mathcal{F}$. Due to Theorem 2.1.3 and the remark below, the following definition is well–posed for fixed p_0.

Definition 2.4.1. *There exists a unique, non–linear, continuous mapping*

$$G : L^2(0,T; L^\infty(\Omega))^m \to W(0,T), \quad u \mapsto G(u), \tag{2.36}$$

such that $p = G(u)$ represents the weak solution of the Fokker–Planck problem (2.2)*–*(2.4)*:*

$$\begin{aligned} &\langle \partial_t p, \cdot \rangle_{H'} + \mathcal{F}[u](p, \cdot) = 0 && \text{in } L^2(0,T; H^1(\Omega)'), \\ &p(0) = p_0 && \text{in } L^2(\Omega). \end{aligned}$$

The operator G maps any admissible control to the associated state and is therefore referred to as the control–to–state operator. In the case of only time–dependent controls, we have an analogous definition for $G : L^2(0,T)^m \to W(0,T)$.

We remark that we prefer to use the same notation G in the case of time–space and only time–dependent controls, i.e., we will write $G : L^2(0,T; L^\infty(\Omega))^m \to W(0,T)$ and $G : L^2(0,T)^m \to W(0,T)$.

Next, we discuss further properties of the control–to–state map G, that is, Fréchet differentiability, Lipschitz continuity and compactness. We will start with a partial result on compactness in the $L^2(\Omega_T)$–norm and then derive differentiability and Lipschitz continuity. With these three properties, we can lastly prove the compactness of G in the $W(0,T)$–norm.

Throughout this chapter, we will often encounter the bilinearity of $\mathcal{F}$ in the control and state argument. What we mean by this is the following. For $u_1, u_2 \in L^2(0,T;L^\infty(\Omega))^m$ and $p_1 := G(u_1), p_2 := G(u_2)$, we have for all test functions $\psi \in H^1(\Omega)$

$$\begin{aligned}\mathcal{F}[u_1](p_1,\psi) - \mathcal{F}[u_2](p_2,\psi) &= \mathcal{F}[u_1]((p_1-p_2),\psi) - \langle p_2\, M(u_1-u_2), \nabla\psi\rangle_{L^2(\Omega)} \\ &= \mathcal{F}[u_2]((p_1-p_2),\psi) - \langle p_1\, M(u_1-u_2), \nabla\psi\rangle_{L^2(\Omega)}.\end{aligned} \tag{2.37}$$

Lemma 2.4.2. *(Compactness of G on the set of admissible controls)*
Let $(u^k)_{k\in\mathbb{N}} \subset \mathcal{U}$, where $\mathcal{U}$ denotes one of the admissible sets under consideration in this chapter, that is

$$\mathcal{U} = U_{\mathrm{ad}}^j \quad \textit{for} \quad j = 1,2,3, \qquad \textit{or} \qquad \mathcal{U} = U_{\mathrm{ad}}^T \;\textit{or}\; U_{\mathrm{ad}}^{T,H}.$$

Then there exists $u \in \mathcal{U}$ such that for a subsequence

$$G(u^k) \to G(u) \;\textit{strongly in}\; L^2(\Omega_T).$$

Proof. Obviously, it is enough to show the assertion for the case $\mathcal{U} = U_{\mathrm{ad}}$ since all the other set of admissible controls can be seen as subset of U_{ad}. Now let $(u^k)_{k\in\mathbb{N}} \subset U_{\mathrm{ad}}$. Due to the box–constraints, each component of u^k is bounded in $L^\infty(\Omega_T)$ uniformly in k by a constant C_{ad}. Hence, there exists $u \in U_{\mathrm{ad}}$ and a weakly* convergent subsequence such that, keeping the same index,

$$u^k \rightharpoonup^* u \quad \text{in } L^\infty(\Omega_T)^m.$$

Notice that $L^\infty(\Omega_T)$ is the dual of $L^1(\Omega_T)$, so we can identify u^k as an element of $L^1(\Omega_T)'$ and obtain that

$$\int_{\Omega_T} u^k(t,x)g(t,x)\,dt\,dx \to \int_{\Omega_T} u(t,x)g(t,x)\,dt\,dx$$

for all $g \in L^1(\Omega_T)$. Next, due to Theorem 2.1.3, we obtain that $G(u^k)$ is bounded in $W(0,T)$ uniformly in k by a constant $C_F C_{\mathrm{ad}} < \infty$. Hence, after possibly extracting a subsequence, $G(u^k)$ converges weakly in $W(0,T)$ to some p, and an application of Corollary 2.1.5 yields that $G(u^k) \to p$ strongly in $L^2(\Omega_T)$ (even strongly in $L^\tau((0,T);L^2(\Omega))$ for all $\tau \geq 1$). Lastly, we need to verify that p solves the FP problem with control u since uniqueness then implies $G(u) = p$. Let us denote $p_k := G(u^k)$ and recall that $a = (a_{ij})$ is the diffusion matrix. Due to the weak convergence of (p_k) in $W(0,T)$, we obtain for the linear terms

$$\begin{aligned}&\int_0^T \left(\langle \dot{p}_k(t), \varphi(t)\rangle_{H'} + \int_\Omega \nabla p_k(t,x)^\top a(t,x)\nabla\varphi(t,x)\,dx\right)dt \\ &\qquad\longrightarrow \int_0^T \left(\langle \dot{p}(t), \varphi(t)\rangle_{H'} + \int_\Omega \nabla p(t,x)^\top a(t,x)\nabla\varphi(t,x)\,dx\right)dt, \quad \text{as } k\to\infty,\end{aligned}$$

for all $\varphi \in H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega))$ with $\varphi(T,\cdot) = 0$. The other linear terms, that is only p_k appears but no u^k, are treated analogously. The interesting part is the convergence of the bilinear term $\int_{\Omega_T} p_k\, Mu^k \cdot \nabla\varphi \to \int_{\Omega_T} p\, Mu \cdot \nabla\varphi$. We include the mixed term $p\, Mu^k \cdot \nabla\varphi$ and observe that

$$\left|\int_{\Omega_T} (p_k - p)(Mu^k)\cdot\nabla\varphi\right| \leq \|p_k - p\|_{L^2(\Omega_T)}\|M\|_{L^\infty(\Omega_T)} C_{\mathrm{ad}} \|\varphi\|_{L^2H^1} \to 0$$

due to the $L^2(\Omega_T)$–strong convergence of $p_k \to p$. Furthermore, it holds that

$$\int_{\Omega_T} p(M(u^k - u))\cdot\nabla\varphi \to 0, \quad \text{as } k\to\infty$$

since $p\,\nabla\varphi^\top M$ is in $L^1(\Omega_T)^m$ and $u^k \rightharpoonup^* u$ in $L^\infty(\Omega_T)^m$. Consequently, passing to the limit, we conclude that p is a weak solution, and therefore, it holds that $p = G(u)$. This completes the proof. □

In the case of only time–dependent controls with controls from $H^1(0,T)^m$, no box–constraints need to be imposed to obtain compactness of G. This is due to compact embedding $L^2(0,T) \Subset H^1(0,T)$ and the a–priori estimate of p w.r.t. u in the $L^2(0,T)$–norm (2.19). The control–to–state map G on $L^2(0,T)^m$ is, however, not compact.

We start with the Fréchet differentiability; notice that it is sufficient to prove this property for the largest space $L^2(0,T;L^\infty(\Omega))^m$, see definition 1.3.3 and the remark below. We consider the functional

$$\begin{aligned} &H : W(0,T) \times L^2(0,T;L^\infty(\Omega))^m \to L^2(0,T;H^1(\Omega)') \times L^2(\Omega), \\ &H(p,u) := \big(H_1(p,u), H_2(p(0))\big) := (\dot{p} + \mathcal{F}[u](p,\cdot), p(0) - p_0). \end{aligned}$$

First, we see that both components of H are arbitrarily often continuously Fréchet differentiable on $W(0,T) \times L^2(0,T;L^\infty(\Omega))^m$. Now, observe that H was defined such that $H(p,u) = (0,0)$ iff p is a solution of the Fokker–Planck problem with drift u and initial PDF p_0. Hence, $H(G(u),u) = (0,0)$ for all $u \in L^2(0,T;L^\infty(\Omega))^m$. Next, we recall the implicit function theorem on Banach spaces. In order to apply this theorem, we have to show that the mapping

$$W(0,T) \ni z \mapsto D_pH(p,u)(z) = (\dot{z} + \mathcal{F}[u](z,\cdot), z(0)) \in L^2(0,T;H^1(\Omega)') \times L^2(\Omega) \tag{2.38}$$

is an isomorphism. This follows immediately from Corollary 2.2.1, specifically, the injectivity follows by the uniqueness and the surjectivity by the existence result.

Hence, the implicit function theorem is applicable for any starting points $(p,u) \in W(0,T) \times L^\infty(\Omega_T)^m$ with $H(p,u) = (0,0)$. Finally, we can deduce that G is continuously Fréchet differentiable in $u \in L^2(0,T;L^\infty(\Omega))^m$ if we apply this theorem in $(G(u),u)$. This yields a continuously Fréchet differentiable function $\tilde{G}$ with $H(\tilde{G}(u),u) = (0,0)$ on an open neighborhood $u \in \tilde{U} \subset L^2(0,T;L^\infty(\Omega))^m$. By uniqueness, $\tilde{G} = G$ on $\tilde{U}$, and since u was chosen arbitrarily, we obtain the differentiability of G on $L^\infty(\Omega_T)^m$.

Furthermore, differentiating $H(G(u),u) = 0$ with respect to u gives an implicit formula for $G'(u)$, namely

$$D_pH(G(u),u)G'(u)(v) + D_uH(G(u),u)(v) = 0, \quad u,v \in L^2(0,T;L^\infty(\Omega))^m. \tag{2.39}$$

Notice that G maps to $W(0,T)$, and hence, for all $u,v \in L^2(0,T;L^\infty(\Omega))^m$, it holds that $G'(u)v \in W(0,T)$. Therefore, we may calculate the Fréchet derivative of H_1 at (p,u) in direction $v \in L^2(0,T;L^\infty(\Omega))^m$. For any test function $\varphi \in W(0,T)$, we have a.e. on $[0,T]$

$$D_uH_1(p,u)(v)(\varphi) = \lim_{\alpha\to 0} \frac{\mathcal{F}[u+\alpha v](p,\varphi) - \mathcal{F}[u](p,\varphi)}{\alpha} = \int_\Omega p\,(Mv)\cdot\nabla\varphi\,dx. \tag{2.40}$$

Plugging (2.40) and (2.38) into (2.39) implies that $z := G'(u)v$ solves the so–called linearized state equation (in weak form) at $(G(u),u) = (p,u)$ in direction $v \in L^\infty(\Omega_T)^m$

$$\begin{aligned} &\langle \dot{z}, \cdot\rangle_{H'} + \mathcal{F}[u](z,\cdot) = \langle f^{\text{lin}}[u,v], \cdot\rangle_{H'} && \text{in } L^2(0,T;H^1(\Omega)'), \\ &z(0) = 0 && \text{a.e. on } \Omega. \end{aligned} \tag{2.41}$$

where for $\psi \in H^1(\Omega)$, $t \in [0,T]$, we define

$$\langle f_t^{\text{lin}}[u,v], \psi\rangle_{H'} := -\sum_{i=1}^{d}\sum_{j=1}^{m} \int_\Omega p(t,x)\, M_{ij}(t,x) v_j(t,x)\, \partial_{x_i}\psi(x)\,dx.$$

The key point is that we can verify that the r.h.s. $\langle f^{\text{lin}}[u,v], \cdot\rangle_{H'}$ is in $L^2(0,T;H^1(\Omega)')$. This follows from the fact that

$$p \in L^\infty(0,T;L^2(\Omega)),\ M \in L^\infty(\Omega_T)^{d\times m} \text{ and } v \in L^2(0,T;L^\infty(\Omega))^m.$$

For the upcoming first– and second–order analysis, it is essential that (2.41) is an inhomogeneous Fokker–Planck problem. Specifically, it is very defining and shaping for our bilinear problem that the source term $f^{\text{lin}}[u,v]$ takes the form from above. Roughly speaking, it is the product of the state and the direction of the derivative, i.e. $f^{\text{lin}}[u,v] \sim G(u)v$.
Now, let us summarize our previous and some further results with the following lemma. We recall the abbreviations for the following Bochner spaces

$$\|\cdot\|_{L^pH^1} := \|\cdot\|_{L^p(0,T;H^1(\Omega))}, \quad \|\cdot\|_{L^pL^q} := \|\cdot\|_{L^p(0,T;L^q(\Omega))}.$$

Lemma 2.4.3. *The control–to–state map G is of class C^∞ in $L^2(0,T;L^\infty(\Omega))^m$. Furthermore, it has the following properties:*

a) Its derivative is the solution of the linearized state equation, i.e., $z := G'(u)v \in W(0,T)$ solves (2.41) for $u, v \in L^2(0,T;L^\infty(\Omega))^m$ and it holds that

$$\|z\|_{L^\infty L^2} + \|z\|_{W(0,T)} \le C_F \|v\|_{L^2L^\infty} \|G(u)\|_{L^\infty L^2}. \tag{2.42}$$

Furthermore, if $u \in L^\infty(\Omega_T)^m$ and $G(u) \in L^\infty(\Omega_T)$, then the following estimate holds

$$\|z\|_{L^\infty L^2} + \|z\|_{W(0,T)} \le C_F \|v\|_{L^2(\Omega_T)} \|G(u)\|_{L^\infty(\Omega_T)}. \tag{2.43}$$

b) G is locally Lipschitz continuous in the following sense:

$$\|G(u) - G(w)\|_{W(0,T)} \le C_F C_1 \|G(w)\|_{L^\infty L^2} \|u - w\|_{L^2L^\infty}, \ u, w \in L^2(0,T;L^\infty(\Omega))^m, \tag{2.44}$$
$$\|G(u) - G(w)\|_{W(0,T)} \le C_F C_2 \|G(w)\|_{L^\infty(\Omega_T)} \|u - w\|_{L^2(\Omega_T)}, \quad u, w \in L^\infty(\Omega_T)^m \tag{2.45}$$

with constants $C_1 = C(\|u\|_{L^2L^\infty})$ and $C_2 = C(\|u\|_{L^\infty(\Omega_T)})$.

c) G is compact in the following sense: If

$$(u^k)_{k\in\mathbb{N}} \subset L^\infty(\Omega_T)^m \textit{ with } u^k \rightharpoonup^* u \textit{ in } L^\infty(\Omega_T)^m$$

or

$$(u^k)_{k\in\mathbb{N}} \subset H^1(0,T)^m \textit{ with } u^k \rightharpoonup u \textit{ in } H^1(0,T)^m,$$

then $G(u^k) \to G(u)$ in $W(0,T)$ strongly.

d) If $d \in \{1,2\}$, then G is also compact in $L^\infty(\Omega_T)$: For

$$(u^k)_{k\in\mathbb{N}} \subset L^\infty(\Omega_T)^m \textit{ with } u^k \rightharpoonup^* u \textit{ in } L^\infty(\Omega_T)^m,$$

it holds that $G(u^k) \to G(u)$ in $L^\infty(\Omega_T)$.

Proof. In order to prove estimate (2.42), we recall that z is a solution of the FP problem with initial state zero and r.h.s. $f^{\text{lin}}[u,v]$. Thus, we can apply Corollary 2.2.1 and observe for $\psi \in H^1(\Omega)$

$$\begin{aligned}\|f^{\text{lin}}[u,v](\psi)\|_{L^2(0,T)} &\le \Big\| \int_\Omega \big|p(\cdot,x)\, \nabla\psi(x)^\top M(\cdot,x)\, v(\cdot,x)\big| \, dx \Big\|_{L^2(0,T)} \\ &\le \|p\|_{L^\infty L^2} \|v\|_{L^2L^\infty} \|M\|_\infty \|\psi\|_{H^1(\Omega)},\end{aligned}$$

and respectively

$$\|f^{\text{lin}}[u,v](\psi)\|_{L^2(0,T)} \le \|p\|_{L^\infty(\Omega_T)} \|v\|_{L^2(\Omega_T)} \|M\|_\infty \|\psi\|_{H^1(\Omega)}.$$

This proves the estimates (2.42) and (2.43). The case of only time–dependent controls, that is $v = v(t) \in L^2(0,T)^m$, is obviously a subproblem of the case $v \in L^2(0,T;L^\infty(\Omega))$, and hence, a bound of p in the $L^\infty(0,T;L^2(\Omega))$–norm is sufficient. Therefore, no bounds of p in the $L^\infty(\Omega_T)$ are necessary in this case. The bound for the derivative of G yields the Lipschitz continuity in the following way: Let u, w be in $L^2(0,T;L^\infty(\Omega))^m$ or $L^\infty(\Omega_T)^m$, and define $z := G(u) - G(w) \in W(0,T)$. Hence, z is a weak solution of the inhomogeneous Fokker–Planck equation with drift u and r.h.s. $f^{\text{lin}}[w, u-w]$. Consequently $z = G'(u)(w-u)$, which is the reason (2.41) is called the linearized state equation. Thus, part a) proves the Lipschitz continuity for controls on $L^2(0,T;L^\infty(\Omega))^m$ or $L^\infty(\Omega_T)^m$. We remark that the estimate (2.45), where the difference $|u-w|$ on the r.h.s. of the Lipschitz–estimate depends on the L^2–norm, turns out to be essential for the second–order analysis of the optimal control problem in the time–space dependent case. This is only possible due to the L^∞–bound of p.

Similarly, we show assertion c) and set $z_k := G(u) - G(u^k) = G'(u)v^k$, $v^k := u - u^k$ for $k \in \mathbb{N}$. According to (2.42), z_k is uniformly bounded in $W(0,T)$, and therefore, there exists some $z \in L^2(0,T;H^1(\Omega_T))$ and $\zeta \in L^2(0,T;H^1(\Omega)')$ such that for a subsequence

$$z_k \rightharpoonup z \quad \text{in } L^2(0,T;H^1(\Omega_T)), \qquad z_k \to z \quad \text{in } L^2(\Omega_T), \qquad \dot{z}_k \rightharpoonup \zeta \quad \text{in } L^2(0,T;H^1(\Omega)'). \tag{2.46}$$

For convenience, we prove that $\dot{z} = \zeta$. Let $\phi \in C_c^\infty(\,]0,T[\,)$ and $\psi \in H^1(\Omega)$, and we interpret the $L^2(0,T;H^1(\Omega))$–function z as $L^2(0,T;H^1(\Omega)')$–function. On the one hand, we have by the weak convergence in $L^2(0,T;H^1(\Omega)')$ that for a subsequence

$$\int_0^T \phi(t)\langle \dot{z}_k(t), \psi\rangle_{H'}\,dt \to \int_0^T \phi(t)\langle \zeta(t), \psi\rangle_{H'}\,dt \quad \text{and} \tag{2.47}$$

$$\int_0^T \dot{\phi}(t)\langle z_k(t), \psi\rangle_{H'}\,dt \to \int_0^T \dot{\phi}(t)\langle z(t), \psi\rangle_{H'}\,dt \quad \text{as } k \to \infty. \tag{2.48}$$

On the other hand, we have for $k \in \mathbb{N}$

$$\int_0^T \phi(t)\langle \dot{z}_k(t), \psi\rangle_{H'}\,dt = -\Big\langle \int_0^T \dot{\phi}(t) z_k(t)\,dt, \psi\Big\rangle_{H'} = -\int_0^T \dot{\phi}(t)\langle z_k(t), \psi\rangle_{H'}\,dt\,; \tag{2.49}$$

the fact that we can interchange the integral and the continuous function $\langle\cdot,\psi\rangle_{H'}$ can be shown straight forwardly by an approximation with simple functions. Since $\psi \in H^1(\Omega)$ was arbitrary, this implies with (2.47) and (2.48) that

$$\int_0^T \phi(t)\zeta(t)\,dt = -\int_0^T \dot{\phi}(t) z(t)\,dt, \quad \text{in } H^1(\Omega)'. \tag{2.50}$$

Finally, $\dot{z} = \zeta$ in $L^2(0,T;H^1(\Omega)')$ follows from the fact that (2.50) holds for every test function $\phi \in C_c^\infty(\,]0,T[\,)$.

Now, we can show that $f^{\text{lin}}[u^k, v^k] \to 0$ in $L^2(0,T;H^1(\Omega)')$, which yields $z_k \to 0$ in $W(0,T)$ according to Corollary 2.2.1. Recall the fact that for any dual $1 < p, q < \infty$ and reflexiv Banach space X, we have that $L^p(0,T;X')$ and $L^q(0,T;X)'$ are isometric isomorph. Hence, for $\varphi \in L^2(0,T;H^1(\Omega))$, it holds that

$$\begin{aligned}\int_0^T f_t^{\text{lin}}[u^k, v^k](\varphi(t))\,dt &= \int_{\Omega_T} G(u^k)(t,x)\,(v^k(t,x)^\top M(t,x))\cdot\nabla\varphi(t,x)\,dt\,dx \\ &\le \int_{\Omega_T} |G(u^k) - G(u)||M||v^k||\nabla\varphi| + \int_{\Omega_T} G(u)\,(Mv^k)\cdot\nabla\varphi,\end{aligned}$$

where we have omitted the (t,x) argument in the second line. We start with the case $u, u^k \in U_{\text{ad}}$, where the weak* convergence holds in $L^\infty(\Omega_T)^m$. Since $\|u^k\|_\infty \le C(1 + \|u\|_\infty)$ for all $k \in \mathbb{N}$ due to the weak* convergence, the first term can be estimated against

$$C_{\|u\|_\infty}\|G(u^k) - G(u)\|^2_{L^2(\Omega_T)}\|\varphi\|^2_{L^2(0,T;H^1(\Omega))},$$

and therefore converges to zero (for a subsequence) due to the compactness result from Lemma 2.4.2. The second term converges to zero since $v^k \rightharpoonup^* 0$ in $(L^1(\Omega_T)^m)'$ and $G(u)\,\nabla\varphi^\top M \in L^1(\Omega_T)^m$.
The case for the weak $H^1(0,T)$–convergence is done analogously, since this implies weak* convergence in $L^\infty(0,T)$. In conclusion, we have shown that $f^{\text{lin}}[u^k, v^k] \to 0$ in $L^2(0,T;H^1(\Omega))'$ after extracting a subsequence.
For part d), we apply Theorem 2.2.3 for

$$z_k = G(u) - G(u^k),\ z_0 = 0, \text{ and r.h.s. } \langle \mathcal{G}, \cdot\rangle_{H'} = \langle f^{\text{lin}}[u,v], \cdot\rangle_{H'}.$$

Consequently, the estimate (2.25) with $g_1 = 0$ and $g_2 = G(u^k)M\,v^k$ implies

$$\|G(u^k) - G(u)\|_{L^\infty(\Omega_T)} \le C(\|G(u^k)M\,v^k\|_{L^q(\Omega_T)} + \|G(u^k) - G(u)\|_{L^q(\Omega_T)}) \tag{2.51}$$

for all $q > 2 + d$. Notice that we have already shown the strong convergence of $G(u^k)$ in $W(0,T)$. Due to Corollary 2.1.5, we have strong convergence in $L^\eta(\Omega_T)$ for all $\eta \le 4/d + 2$. Since $d \in \{1,2\}$ we can conclude that both $L^q(\Omega_T)$–norms in (2.51) tend to zero, and therefore, $G(u^k)$ converges to $G(u)$ in $L^\infty(\Omega_T)$ as desired.
Moreover, the above can be applied to any subsequence of the original sequence. Thus, every subsequence of $G(u^k)$ has a sub–subsequence converging to the same limit $G(u)$ since weak solutions to the inhomogeneous FP problem are unique. Consequently, in c) and d) we do not need a selection of a subsequence after an application of Lemma 1.4.3. This concludes the proof. □

With the same techniques, we obtain the weak formulation of the second–order Fréchet derivative of G. Let $u \in U_{\text{ad}}$ and $v_1, v_2 \in L^\infty(0,T)^m$. Then, the function $w := G''(u)(v_1,v_2) \in W(0,T)$ satisfies

$$\begin{aligned} &\langle \dot{w}, \cdot\rangle_{H'} + \mathcal{F}[u](w,\cdot) = \langle f^{\text{quad}}, \cdot\rangle_{H'} && \text{in } L^2(0,T;H^1(\Omega)'), \\ &w(0) = 0 && \text{a.e. on } \Omega, \end{aligned} \tag{2.52}$$

where for $\psi \in H^1(\Omega)$, $t \in [0,T]$, we define the r.h.s

$$\langle f_t^{\text{quad}}[u,v_1,v_2], \psi\rangle_{H'} := -\int_\Omega \big(z_1(t,x)\,v_2(t,x)^\top M(x) + z_2(t,x)\,v_1(t,x)^\top M(x)\big)\cdot\nabla\psi(x)\,dx.$$

Due to Corollary 2.2.1 on inhomogeneous FP problems, we obtain analogous estimates for w in the $W(0,T)$–norm as in Lemma 2.4.3 a), that is

$$\|w\|_{L^\infty(0,T;L^2(\Omega))} + \|w\|_{L^2(0,T;H^1(\Omega))} + \|\dot{w}\|_{L^2(0,T;H^1(\Omega)')} \le C_{\text{F}}C_{\text{ad}}\|f^{\text{quad}}[u,v_1,v_2]\|_{L^2(0,T;H^1(\Omega)')}.$$

Moreover, we obtain similar results for the Fréchet derivatives of G. Since these results are only used for the numerical analysis of the Galerkin discretization presented in Chapter 5 and 7, we prove it only for time–dependent controls.

Lemma 2.4.4. *G' and G'' are compact in the following sense: If $v \in L^2(0,T)^m$ and $(u^k)_{k\in\mathbb{N}} \subset U_{\text{ad}}^T$ with $u^k \rightharpoonup^* u$ in $L^\infty(0,T)^m$ as $k \to \infty$, then*

$$G'(u^k)v \to G'(u)v, \quad G''(u^k)(v,v) \to G''(u)(v,v) \qquad \text{in } L^\infty(0,T;L^2(\Omega)), \quad \text{as } k \to \infty.$$

Furthermore, G' and G'' are Lipschitz continuous, globally on U_{ad}^T, in the sense that there exists a constant $C = C_{\text{ad}}C_{\text{F}} > 0$ such that for all $u, w \in U_{\text{ad}}^T$, $v \in L^2(0,T)^m$

$$\|G'(u)v - G'(w)v\|_{L^\infty L^2} \le C\|u-w\|_2\|v\|_2 \tag{2.53}$$

$$\|G''(u)(v,v) - G''(w)(v,v)\|_{L^\infty L^2} \le C\|u-w\|_2\|v\|_2^2. \tag{2.54}$$

Proof. First, we want to prove the compactness. Let $p := G(u)$, $p_k := G(u^k)$, $z_k := G'(u^k)v$ and $z := G'(u)v$, and define $\delta z := z - z_k$, $\delta p = p - p_k$, $\delta u = u - u^k$. Notice that $(z_k)_{k\in\mathbb{N}}$ is uniformly bounded in $W(0,T)$, and therefore, it possesses a subsequence (denoted in the same way) which converges strongly in $L^2(\Omega_T)$. We refer to (2.46) for the proof of this assertion. Next, due to the bilinear structure of $\mathcal{F}$, see (2.37), it holds that

$$\mathcal{F}[u](z,\psi) - \mathcal{F}[u^k](z_k,\psi) = \mathcal{F}[u](\delta z,\psi) + \langle z_k\, M\delta u, \nabla\psi\rangle_{L^2(\Omega)}, \quad \psi \in H^1(\Omega).$$

Furthermore, we have $f^{\mathrm{lin}}[u,v] - f^{\mathrm{lin}}[u^k,v] = \langle \delta p\, Mv, \nabla\cdot\rangle_{L^2(\Omega)}$ in the $L^2(0,T;H^1(\Omega)')$–sense. Therefore, δz solves the inhomogeneous problem a.e. on $[0,T]$

$$\begin{aligned}\langle \dot{\delta z}, \psi\rangle_{H'} + \mathcal{F}[u](\delta z,\psi) &= \langle \delta p\, Mv - z_k\, M\delta u, \nabla\psi\rangle_{L^2(\Omega)}, \quad \psi \in H^1(\Omega),\\ \delta z(0) &= 0 \quad \text{a.e. on } \Omega.\end{aligned}$$

For all $\psi \in H^1(\Omega)$, let us define

$$\mathcal{G}(\psi) := \langle \delta p\, Mv - z_k\, M\delta u, \nabla\psi\rangle_{L^2(\Omega)}$$

on $[0,T]$. Then, $\mathcal{G} \in L^2(0,T;H^1(\Omega)')$, and we can apply the estimate on inhomogeneous FP problems from Corollary 2.2.1. Thus, we obtain the following bound

$$\|\delta z\|_{L^\infty L^2} + \|\delta z\|_{W(0,T)} \le C_{\mathrm{ad}} C_{\mathrm{F}} \|\mathcal{G}\|_{L^2(0,T;H^1(\Omega)')}. \tag{2.55}$$

Next, we show that $\|\mathcal{G}\|_{L^2(0,T;H^1(\Omega)')}$ tends to zero as k tends to infinity. First, recall that $L^2(0,T;H^1(\Omega)')$ and $L^2(0,T;H^1(\Omega))'$ are isometric isomorph. Next, we exploit the compactness of G to find that $\delta p \to 0$ strongly in $L^\infty(0,T;L^2(\Omega))$. Thus, it holds for all $\varphi \in L^2(0,T;H^1(\Omega))$ that

$$\int_{\Omega_T} \delta p\, \nabla\varphi^\top M\, v\, dt\, dx \le \|\delta p\|_{L^\infty L^2}\|v\|_2\|M\|_\infty\|\nabla\varphi\|_{L^2(\Omega_T)} \le C_{\mathrm{F}}\|\delta p\|_{L^\infty L^2}\|v\|_2\|\varphi\|_{L^2H^1} \to 0$$

as $k \to \infty$. Let us consider the second term in $\mathcal{G}$. Since $\delta u \rightharpoonup^* 0$ in $L^\infty(0,T)^m$, we observe that the strong $L^2(\Omega_T)$–convergence of z_k (for a subsequence) yields the convergence

$$\int_{\Omega_T} z_k\, \nabla\varphi^\top M\, \delta u\, dt\, dx \to 0, \quad \varphi \in L^2(0,T;H^1(\Omega))$$

as $k \to \infty$ for a subsequence. For the same subsequence, this implies the convergence of the r.h.s.

$$\|\mathcal{G}\|_{L^2(0,T;H^1(\Omega)')} \to 0, \qquad \text{as } k \to \infty,$$

and therefore, $G'(u^k)v$ converges to $G'(u)v$ in the desired norm for a subsequence. Since $G'(u)v$ is the unique solution, an application of Lemma 1.4.3 implies that the convergence of $G'(u^k)v$ holds even without selecting a subsequence, and we have proven the compactness of G'. Once this result is established, the proof for the compactness of G'' can be done analogously.

The proof for the Lipschitz continuity is done similarly, where obviously u^k has to be replaced by w in the definitions from above. Thus, we arrive at the same estimate (2.55) for δz. Now, we exploit the Lipschitz continuity of G and estimate $\|\delta p\|_{L^\infty L^2} \le C_{\mathrm{ad}}C_{\mathrm{F}}\|u-w\|_2$, which implies

$$\int_{\Omega_T} \delta p\, \nabla\varphi^\top M\, v\, dt\, dx \le C_{\mathrm{ad}}C_{\mathrm{F}}\|u-w\|_2\|v\|_2\|\varphi\|_{L^2H^1}, \quad \varphi \in L^2(0,T;H^1(\Omega)).$$

Since $\|z_k\|_{W(0,T)} \le C_{\mathrm{ad}}C_{\mathrm{F}}\|v\|_2$, we observe that

$$\int_{\Omega_T} z_k\, \nabla\varphi^\top M\, \delta u\, dt\, dx \le C_{\mathrm{ad}}C_{\mathrm{F}}\|u-w\|_2\|v\|_2\|\varphi\|_{L^2H^1}, \quad \varphi \in L^2(0,T;H^1(\Omega)).$$

Finally, combining both estimates for the two terms in $\mathcal{G}$ gives us the desired estimate

$$\|\mathcal{G}\|_{L^2(0,T;H^1(\Omega)')} \leq C_{\mathrm{ad}} C_{\mathrm{F}} \|u - w\|_2 \|v\|_2.$$

This concludes the proof of the Lipschitz continuity of G'. Once the Lipschitz continuity of G' is shown, the desired Lipschitz estimate for G'' can be proven completely analogously. □

3

Ensemble optimal control problems governed by the Fokker–Planck equation

Einstein had, for the first time connected new and measurable consequences to statistical physics. That might sound like a largely technical achievement, but on the contrary, it represented the triumph of a great principle: that much of the order we perceive in nature belies an invisible underlying disorder and hence can be understood only through the rules of randomness.

Leonard Mlodinow in The Drunkard's Walk: How Randomness Rules Our Lives, 2008

In this chapter, we analyze the ensemble optimal problem that has been derived in Section 1.2. The optimization problem under consideration, in its most general form, reads

$$\begin{aligned}&\min_{u\in\mathcal{U}} J(p,u) \quad p \text{ subject to}\\ &\partial_t p + \mathcal{F}[u](p,\cdot) = 0 \quad \text{in } L^2(0,T;H^1(\Omega)'),\\ &p(0) = p_0 \quad \text{a.e. on } L^2(\Omega),\end{aligned}$$

where

$$J(p,u) := \int_{\Omega_T} \mathcal{R}[u](t,x)p(t,x)\,dt\,dx + \int_\Omega \mathcal{T}(x)\,p(T,x)\,dx + \frac{\gamma_2}{2}\|u\|_Y^2, \tag{3.1}$$

$$\text{and} \quad \mathcal{R}[u](t,x) := \frac{\gamma_1}{2}|u(t,x)|^2 + \alpha(t,x)\cdot u(t,x) + \beta(t,x). \tag{3.2}$$

The bilinear form $\mathcal{F}[u](\cdot,\cdot)$ can be found in (2.8). Using the control–to–state map and defining $\hat{J}(u) := J(G(u),u)$ for all u from the admissible set $\mathcal{U}$, the optimal control problem is reformulated as the minimization problem

$$\min_{u \in \mathcal{U}} \hat{J}(u). \tag{3.3}$$

In this thesis, the main focus for $\mathcal{U}$ is put on box–constrained controls.

In the framework of ensemble control problems when multiplied with a density function, $\mathcal{R}[u]$ and $\mathcal{T}$ are referred to as running cost and terminal cost, respectively. The third term in (3.1) is not subject to averaging, and therefore, it does not belong to a typical formulation of an ensemble problem. However, for the theoretical analysis of (3.3), it will be necessary in some cases to assume $\gamma_2 > 0$, as it will imply very useful properties to solutions $\bar{u}$ of (3.3). One property will be that such $\bar{u}$ has higher regularity than one expects, and hence, this parameter γ_2 is often referred to as regularization parameter. The parameter γ_1 has a similar effect on solutions $\bar{u}$ under a strict positivity assumption on the PDF p and is referred to as the quadratic cost term.

Throughout this chapter, we impose the following natural assumptions on these quantities:

(J1)

$$\alpha \in L^\infty(\Omega_T)^m, \quad \beta \in L^\infty(\Omega_T), \quad \mathcal{T} \in L^\infty(\Omega) \cap H^1(\Omega),$$

(J2)

$$\gamma_1, \gamma_2 \geq 0.$$

We denote with $C_J > 0$ a generic constant that depends continuously on the quantities from (J1) in the corresponding norms.

The choices for the set of admissible controls $\mathcal{U}$ and the norm of the regularizing term Y that we consider are the following:

(U1) For time–space dependent controls, we recall

$$Y_1 = L^2(\Omega_T)^m, \quad Y_2 = L^2(0,T;H_0^1(\Omega))^m, \quad Y_3 = H^1(\Omega_T)^m$$

and the set of admissible controls $U_{\text{ad}}^j = U_{\text{ad}} \cap Y_j$ for $j \in \{1,2,3\}$.

(U2) For only time–dependent controls we have

$$Y_T = L^2(0,T)^m \text{ and } Y_H = H_0^1(0,T)^m,$$

where the sets U_{ad} and U_{ad}^T from (2.35) for given $-\infty < u^{\min} < u^{\max} < \infty$ are as follows

$$\begin{aligned} U_{\text{ad}} &= \left\{ u \in L^\infty(\Omega_T)^m \,:\, u^{\min} \leq u_i \leq u^{\max}, \quad \text{a.e. on } \Omega_T,\ i = 1,\dots,m \right\}, \\ U_{\text{ad}}^T &= \left\{ u \in L^\infty(0,T)^m \,:\, u^{\min} \leq u_i \leq u^{\max}, \quad \text{a.e. on } [0,T],\ i = 1,\dots,m \right\}. \end{aligned}$$

The questions about problem (3.3) that we mainly investigate are existence of optimal controls, the well–posedness of the adjoint problem, implicit equations and higher regularity for local minima, uniqueness, and coercivity. Obviously, different settings for the set of admissible controls and regularizing norm will result in different outcomes and properties of the problem. In the next section, we want to give an overview on this.

3.1 Main results – an overview

The minimum requirement for the analysis of any optimal control problem is – obviously – that solutions exist, and therefore, we start with this issue. In a nutshell, the bottleneck to derive existence of optimal controls for our ensemble optimal control problem are the following two criteria: The boundedness of $\hat{J}$ from below in the sense that

$$\inf_{u\in\mathcal{U}} \hat{J}(u) > -\infty$$

and the weak lower semi–continuity (w.l.s.c) of $\hat{J}$ on $\mathcal{U}$, where $\mathcal{U}$ denotes one of the set of admissible controls from above. More precisely, the w.l.s.c. of $\hat{J}$ follows if the control–to–state map is compact in the sense that the set $\{G(u) \mid u \in \mathcal{U}\}$ is relatively compact in $L^2(\Omega_T)$. Therefore, the existence of optimal controls depends on the choice of the admissible set $\mathcal{U}$.

On the one hand, Lemma 2.4.3 gives us a criterion on $\mathcal{U}$ for the compactness of G. On the other hand, finding a uniform lower bound of $\hat{J}$ is very troublesome due to the term $\alpha \cdot u\,G(u)$, even if we put restrictions to the function α. Since the Fokker–Planck problem is a bilinear problem in $(G(u), u)$, a bound of $\|G(u)\|_{L^2(\Omega_T)}$ does, in general, depend non–linearly on u in an adequate norm. Therefore, both terms $\alpha\cdot u\,G(u)$ and $\beta\,G(u)$, possibly tending to $-\infty$, cannot be compensated by the quadratic term $\gamma_2\|u\|_Y^2$, and one cannot find a lower bound of $\hat{J}$ on $\mathcal{U}$ without the presence of box–constraints, in general.

It is possible to obtain existence of optimal controls with no box–constraints if $\gamma_2 > 0$ and $\alpha = 0$, $\beta \geq 0$, since this obviously implies that $\hat{J}$ is non–negative; notice that p is a PDF and therefore non–negative a.e. on Ω_T.

In conclusion, in the case of an ensemble optimal control problem, we can only prove existence of optimal controls when box constraints are present, or when a lower box–constraint $u^{\min} \geq 0$ is active with additional assumption $\alpha, \beta \geq 0$ on Ω_T.

Let us mention that when considering any stronger regularizing norm, say $Y = L^2(0,T;H^s(\Omega))$ where $s > d/2$, one would still need box constraints for the existence of optimal controls for the same reasoning. The embedding $L^\infty(\Omega) \subset H^s(\Omega)$ would yield sufficient compactness of G; however, it is in general not possible to prove

$$\inf_{u\in L^2(0,T;H^s(\Omega))} \hat{J}(u) > -\infty.$$

Next, we consider the adjoint problem that will be derived in Section 3.3 below. The approach via the adjoint will be our essential tool for a first– and second–order analysis of (3.3), and hence, existence of sufficient regular solutions is necessary. The classical formulation of the adjoint problem for q with corresponding control u reads

$$\begin{aligned} \partial_t q + L^* q &= -\mathcal{R}[u] && \text{on } \Omega_T, \\ q(T) &= \mathcal{T} && \text{on } \Omega, \\ \nabla q \cdot \hat{n} &= 0 && \text{on }]0,T[\times\partial\Omega. \end{aligned}$$

We remark that the PDF $p = G(u)$ does not appear in the adjoint formulation, since J is affine linear in p.

In general, we are only able to prove that $u \in L^2(0,T;L^\infty(\Omega))$ implies the existence of distributional solutions q merely in $L^2(0,T;H^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$, whereas $u \in L^\infty(\Omega_T)$ is sufficient for the existence of weak solutions $q \in W(0,T)$. In this case, any L^2–local minimizer $\bar{u}$ satisfies the following variational inequality

$$\langle(\gamma_1\bar{u} + \alpha - M\nabla\bar{q})\bar{p}, v - \bar{u}\rangle_{L^2(\Omega_T)} + \gamma_2\langle\bar{u}, v - \bar{u}\rangle_Y \geq 0, \quad v \in \mathcal{U},$$

where $\bar{p}$ and $\bar{q}$ are the state and adjoint corresponding to the control $\bar{u}$. This inequality is also referred to as the optimality condition, and $(\bar{u}, \bar{p}, \bar{q})$ is the optimality triplet.

Let us illustrate how to obtain higher regularity from the optimality condition. We take $\gamma_1 = 0$, $\gamma_2 > 0$ and $\mathcal{U} = U_{\mathrm{ad}}$, $Y = L^2(\Omega_T)$ as an example. Then, the variational inequality becomes

$$\langle (\alpha - M\nabla q)p + \gamma_2 u, v - u\rangle_{L^2(\Omega_T)} \geq 0, \quad v \in U_{\mathrm{ad}},$$

which implies the implicit representation

$$\bar{u} = \frac{1}{\gamma_2} \min\left\{u^{\max}, \max\left\{u^{\min}, (\alpha - M\nabla \bar{q})\bar{p}\right\}\right\} \quad \text{a.e. on } \Omega_T.$$

Assuming α and M are smooth, this implies that $\bar{u}$ obtains (up to a certain degree) the regularity of $\nabla\bar{q}$ and $\bar{p}$. On the other hand, higher regularity of $\bar{u}$ may yield higher regularity of $\bar{p}$ and $\bar{q}$. Consequently, this bootstrap argument leads in some cases to higher regularity of the triplet $(\bar{u}, \bar{p}, \bar{q})$. In conclusion, investigating the adjoint problem, the first derivative of $\hat{J}$, the optimality condition and deriving higher regularity of optimal controls is mostly what is considered to be a first–order analysis of the optimization problem (3.3).

The next step is a second–order analysis of this optimal control problem. Since the control–to–state map is non–linear, the functional to be minimized $\hat{J}$ is non–convex. Therefore, a second–order analysis has to be added to a first–order analysis to investigate the following questions:

- If a triplet $(\bar{u}, \bar{p}, \bar{q})$ satisfies the optimality condition, under which further condition is $\bar{u}$ a local and strict minimum of $\hat{J}$?
- Under what conditions are minimizers isolated, i.e., there are no critical points nearby?
- Are local minimizers $\bar{u}$ stable, in the sense that small changes of functions in $\hat{J}$ only lead to small changes of $\bar{u}$?

Furthermore, a second–order analysis for non–convex problems is essential for its numerical analysis. For instance, the convergence analysis of numerical methods are usually based on second–order sufficient optimality conditions. We will see that these questions can be answered by Theorem 1.3.5, and therefore, it will be our main concern to prove that the reduced cost functional $\hat{J}$ satisfies the conditions (C1) and (C2).

From here on, regarding a second–order analysis with the theory presented in Section 1.3, we will restrict ourselves further to the case $\mathcal{U} = U^2_{\mathrm{ad}}$ and regularizing norm $Y = H^1(\Omega_T)$ with $\gamma_2 > 0$. More precisely, it seems difficult to verify condition (C2) in a more general setting. Particularly, in order to prove (C2.2), it seems necessary to consider controls with the additional H^1–regularity and the choice of the regularizing term $\|\cdot\|_Y$ has to include these derivatives. Therefore, we want to apply Theorem 1.3.5 with the space $U_2 = H^1(\Omega_T)^m$ instead of the canonical space $U_2 = L^2(\Omega_T)^m$.

Next, we give an overview for the case of only time–dependent controls $u = u(t)$. Obviously, this is the easier problem in the sense that all the results from the time–space dependent case hold true in an analogous setting for time–dependent controls. In some cases, we have additional freedom in the choice of $(\mathcal{U}, Y)$ to obtain similar results for a first– and second–order analysis.

Firstly, notice that the quadratic cost term becomes obsolete in the case of only time–dependent controls, since p is a PDF due to assumption (F3) and Corollary 2.1.4. Thus, it holds that

$$\frac{\gamma_1}{2}\int_{\Omega_T} |u(t)|^2 p(t,x)\,dt\,dx = \frac{\gamma_1}{2}\int_0^T |u(t)|^2 \int_\Omega p(t,x)\,dx\,dt = \frac{\gamma_1}{2}\|u\|^2_{L^2(0,T)},$$

and we may assume that $\gamma_1 = 0$. The different settings for $\mathcal{U}$ and γ_2 that we consider, and in which the existence of optimal controls is ensured, are the following

$$\begin{aligned} &\mathcal{U} = U^T_{\mathrm{ad}} && \text{if } \gamma_2 \geq 0, \\ \text{or } &\mathcal{U} = H^1(0,T)^m \text{ or } H^1_0(0,T)^m && \text{if } \gamma_2 > 0, \end{aligned}$$

and $Y = L^2(0,T)^m$ or $H^1(0,T)^m$.
We obtain an analogous adjoint problem that has weak solutions in $W(0,T)$ if

$$\mathcal{U} = U_{\text{ad}}^T,\ H^1(0,T)^m \text{ or } H_0^1(0,T)^m.$$

For these admissible sets $\mathcal{U}$, we obtain for any L^2–local minimizer $\bar{u}$ the following optimality condition for the triplet $(\bar{u}, \bar{p}, \bar{q})$

$$\langle \Phi[\bar{u}], v - \bar{u}\rangle_{L^2(0,T)} \ + \gamma_2 \langle \bar{u}, v - \bar{u}\rangle_Y \geq 0, \quad v \in \mathcal{U},$$

where we define for $t \in [0,T]$

$$\Phi[\bar{u}](t) := \int_\Omega \big(\alpha(t,x) - M(t,x)\nabla \bar{q}(t,x)\big)\bar{p}(t,x)\,dx.$$

3.2 Existence of optimal controls

In this section, we prove existence of optimal controls of (3.3) under the different settings for $(\mathcal{U}, Y)$ from (U1) and (U2). A fundamental property to derive existence of minimizers is the weak lower semi–continuity of $\hat{J}$. This property follows, more or less, from the weak lower semi–continuity of $J(p,\cdot)$ for fixed $p \in W(0,T)$ in the second argument, and from the compactness of G. We recall that throughout this chapter, the assumptions (F1)–(F3) and (J1)–(J2) hold.

Lemma 3.2.1. *The reduced cost functional $\hat{J} : U \to \mathbb{R}$ is w.l.s.c. for $U = L^\infty(\Omega_T)$ in the case of time–space dependent controls, or $U = L^\infty(0,T)^m, H^1(0,T)^m$ or $H_0^1(0,T)^m$ in the case of only time–dependent controls.*

Proof. By the same reasoning as in the proof of Lemma 2.4.2, it is enough to consider the case $U = L^\infty(\Omega_T)^m$. Let $(u^k) \subset L^\infty(\Omega_T)^m$ and $u \in L^\infty(\Omega_T)^m$ with $u^k \rightharpoonup^* u$ in $L^\infty(\Omega_T)$. Let $p_k := G(u^k)$ and $p := G(u)$. Recall that

$$\begin{aligned}\hat{J}(u^k) = &\int_{\Omega_T} \Big(\frac{\gamma_1}{2}|u^k(t,x)|^2 + \alpha(t,x)\cdot u^k(t,x) + \beta(t,x)\Big)p_k(t,x)\,dt\,dx \\ &+ \int_\Omega \mathcal{T}(x)p_k(T,x)\,dx + \frac{\gamma_2}{2}\|u^k\|_Y^2.\end{aligned}$$

Let us go through the convergence of each term in $\hat{J}(u^k)$: For any choice of Y

$$Y = L^2(\Omega_T),\ L^2(0,T;H_0^1(\Omega)) \text{ or } H^1(\Omega_T)$$

we know that the Y–norm is w.l.s.c., that is, $\|u\|_Y \leq \liminf_{k\to\infty} \|u^k\|_Y$. Next, we apply Lemma 2.4.3 c) to derive that $p_k \to p$ in $W(0,T)$ strongly. For the linear terms, we obviously have

$$\int_\Omega \mathcal{T}(x)p_k(T,x)\,dx \to \int_\Omega \mathcal{T}(x)p(T,x)\,dx \quad \text{and} \quad \int_{\Omega_T} \beta(t,x)p_k(t,x) \to \int_{\Omega_T} \beta(t,x)p(t,x),$$

due to the convergence in $L^\infty(0,T;L^2(\Omega))$ and $L^1(\Omega_T)$. We can conclude the convergence of the bilinear term after adding a mixed term, that is,

$$\int_{\Omega_T} \alpha\cdot(u^k p_k - up) = \int_{\Omega_T} \alpha\cdot u^k(p_k - p) + \int_{\Omega_T} \alpha\cdot(u^k - u)p \to 0.$$

Lastly, the convergence

$$\int_{\Omega_T} |u|^2 p \leq \liminf_{k\to\infty} \int_{\Omega_T} |u^k|^2 p_k$$

follows from an application of Lemma 1.4.1. For that purpose, we exploit the non–negativity of p_k and p, the strong convergence $p_k \to p$ in $L^1(\Omega_T)$ and the weak* convergence of u^k in $L^\infty(\Omega_T)$. This concludes the proof. □

Theorem 3.2.2. *(Existence of optimal controls)*
Let the conditions (F1)–(F3) and (J1)–(J2) hold. The optimal control problem

$$\min_{u\in\mathcal{U}} \hat{J}(u) \tag{3.4}$$

possesses at least one solution $\bar{u}$ for $\mathcal{U}$ from (U1) or (U2).

Proof. We only consider the case of time–space dependent controls, the case of only time–dependent controls is shown analogously. The first step is to show that $\hat{J}$ is bounded from below on the set $\mathcal{U}$. Due to the box–constraints, every control $u \in \mathcal{U}$ is bounded in the $L^\infty(\Omega_T)$–norm by $C_{\mathrm{ad}} > 0$. This fact, together with Theorem 2.1.3 b), implies that the set of corresponding states $\{G(u) \mid u \in \mathcal{U}\}$ is bounded in $W(0,T)$. Consequently, we obtain the boundedness of $\hat{J}$ from below, that is, for every $u \in \mathcal{U}$ it holds that

$$\begin{aligned}\hat{J}(u) = &\int_{\Omega_T} \Big(\frac{\gamma_1}{2}|u(t,x)|^2 + \alpha(t,x)\cdot u(t,x) + \beta(t,x)\Big) p(t,x)\, dt\, dx \\ &+ \int_\Omega \mathcal{T}(x) p(T,x)\, dx + \frac{\gamma_2}{2}\|u\|_Y^2 \\ \geq & - C_F C_J C_{\mathrm{ad}} > -\infty.\end{aligned}$$

Hence, there exists a minimizing sequence denoted by $(u^k) \subset \mathcal{U}$ such that

$$\hat{J}(u^k) \to I := \inf_{u\in\mathcal{U}} \hat{J}(u).$$

This sequence of minimizers converges (after extracting a subsequence) weakly* to some $\bar{u} \in \mathcal{U}$ in $L^\infty(\Omega_T)^m$, and hence, Lemma 3.2.1 implies that

$$I \leq \hat{J}(\bar{u}) \leq \liminf_{k\to\infty} \hat{J}(u^k) = I.$$

This implies that $\bar{u}$ is a minimizer of $\hat{J}$.

□

We can conclude this section by recalling the definition of a local minimizer of (3.4); compare this to Definition 1.3.4.

Definition 3.2.3. *(Local minimizers)*
Let $\bar{u} \in U_{\mathrm{ad}}$ and let

$$\|\cdot\|_Y = \|\cdot\|_{L^2(\Omega_T)},\ \|\cdot\|_{L^2(0,T;H^1(\Omega))} \ \text{or}\ \|\cdot\|_{H^1(\Omega_T)}.$$

We say that $\bar{u}$ is a Y–local solution of (1.22) *or a local minimizer of J in Y (or w.r.t. the Y–norm), if there exists some $\varepsilon > 0$ such that $J(\bar{u}) \leq J(u)$ holds for all $u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; Y)$. If $J(\bar{u}) < J(u)$ holds for this set with $u \neq \bar{u}$, we say that $\bar{u}$ is a strict minimizer in Y and locally unique in Y (or w.r.t. the Y–norm). If we state that $\bar{u}$ is a local minimizer, this means that it is a local minimizer w.r.t. the $L^\infty(\Omega_T)$–norm.*

Obviously, we have an analogous definition for only time–dependent controls. Notice that the following implications hold: If $\bar{u}$ is a $L^2(\Omega_T)$–local solution, then it is a local solution, and if $\bar{u}$ is a $L^2(0,T;H^1(\Omega_T))$–local solution, then it is a $H^1(\Omega_T)$–local solution.
Furthermore, every Y–local solution $\bar{u}$ is characterized by the following first–order necessary optimality condition

$$\hat{J}'(\bar{u})(u - \bar{u}) \geq 0, \quad u \in U_{\mathrm{ad}} \cap Y.$$

It turns out that $\hat{J}'(\bar{u})$ can be represented in a very useful way by the following adjoint problem that is discussed next.

3.3 The adjoint problem

In order to motivate the need of an adjoint problem, let us calculate the Fréchet derivative of $\hat{J}$. Let $(\mathcal{U}, Y) = (U^j_{\text{ad}}, Y_j)$ for $j = 1, 2, 3$ denote the set of admissible controls with corresponding regularizing norm $\| \cdot \|_Y$, where we focus only on time–space dependent controls. By an application of the chain rule for $\hat{J}(\cdot) = J(G(\cdot), \cdot)$, we obtain for any $u, v \in \mathcal{U}$ with $p := G(u)$ and $z := G'(u)v$ the following equation

$$J'(u)v = \int_{\Omega_T} p\,(\gamma_1 u + \alpha) \cdot v\, dt\, dx + \int_{\Omega_T} \mathcal{R}[u] z\, dt\, dx + \int_{\Omega} \mathcal{T}\, z(T)\, dx + \gamma_2 \langle u, v \rangle_Y.$$

First, we see that the terms are well–defined for $v \in Y$. Furthermore, excluding the last term for the moment, $v \mapsto \hat{J}'(u)v$ is a linear and bounded mapping on $L^2(\Omega_T)^m$, and consequently there has to exist some function $\Phi \in L^2(\Omega_T)$ such that $\hat{J}'(\bar{u}) = \Phi[\bar{u}] + \gamma_2 \bar{u}$ in the sense that

$$\hat{J}'(\bar{u})v = \langle \Phi[\bar{u}], v \rangle_{L^2(\Omega_T)} + \gamma_2 \langle \bar{u}, v \rangle_Y, \quad v \in \mathcal{U}.$$

Obviously, such a representation allows us to obtain an implicit formula for $\bar{u}$, and it is our goal to determine the function Φ. For that purpose, we will rewrite the terms $\int_{\Omega_T} \mathcal{R}[u]z$ and $\int_{\Omega} \mathcal{T}\, z(T)$ by using the adjoint state q, defined next.

Definition 3.3.1. *For any control $u \in L^\infty(\Omega_T)^m$, we say that $q \in W(0, T)$ is the weak solution of the adjoint problem, or q is the adjoint, if f.a.e. $t \in [0, T]$ and for all $\psi \in H^1(\Omega)$ it holds that*

$$-\langle \dot{q}(t), \psi \rangle_{H'} + \mathcal{F}_t[u](\psi, q(t)) = \langle \mathcal{R}[u](t), \psi \rangle_{L^2(\Omega)}, \tag{3.5}$$

$$q(T) = \mathcal{T}, \quad \textit{a.e. on } \Omega. \tag{3.6}$$

We recall that the assumptions (F1)–(F3) and (J1)–(J2) hold, and the adjoint of the FP operator, see (2.5), reads

$$L^* q = \sum_{i,j=1}^{d} a_{ij}(t, x) \partial^2_{x_i x_j} q + \sum_{i=1}^{d} b_i(t, x) \partial_{x_i} q.$$

The classical formulation for the adjoint problem is the following backward problem

$$\begin{aligned} -\partial_t q - L^* q &= \mathcal{R}[u] && \text{on } \Omega_T, \\ q(T) &= \mathcal{T} && \text{on } \Omega, \\ \nabla q \cdot \hat{n} &= 0 && \text{on }]0, T[\, \times \partial\Omega. \end{aligned}$$

Theorem 3.3.2. *a) For every $u \in L^\infty(\Omega_T)^m$, there exists a unique solution $q \in W(0, T)$ of (3.5)–(3.6), and it satisfies the estimate*

$$\|q\|_{W(0,T)} \le C_F C_u \big(\|\mathcal{T}\|_2 + \|\mathcal{R}[u]\|_{L^2(0,T;H^1(\Omega)')} \big) \le C_F C_u C_J, \tag{3.7}$$

where $C_u > 0$ depends continuously on $\|u\|_{L^\infty(\Omega_T)}$. The mapping

$$\Theta : L^\infty(\Omega_T)^m \to W(0, T), \quad u \mapsto \Theta(u) = q$$

is well–defined and referred to as control–to–adjoint map.

b) If additionally (F5)–(F7) holds, then q is bounded in $L^2(0, T; H^2(\Omega)) \cap H^1(0, T; L^2(0, T))$ by a constant $C = C_{\text{ad}} C_{\text{F}*} C_J C_u$

$$\|q\|_{L^2(0,T;H^2(\Omega))} + \|q\|_{H^1(0,T;L^2(\Omega))} < C. \tag{3.8}$$

c) If $d \in \{1,2\}$ additionally to b), then q is also essentially bounded and there exists a constant $C = C_{\text{ad}} C_{\text{F}*} C_J C_u$ such that

$$q\|_{L^\infty(\Omega_T)} < C,$$

where C_u depends continuously only on $\|u\|_\infty$.

Proof. The existence of weak solutions in $W(0,T)$ follows similarly as the a priori estimates for the Fokker–Planck problem after the time transformation $t \mapsto T - t$; notice that we have for the r.h.s.

$$\|\mathcal{R}[u]\|^2_{L^2(0,T;H^1(\Omega)')} \leq \int_0^T \left\| \frac{\gamma_1}{2} |u(t)|^2 + \alpha(t) \cdot u(t) + \beta(t) \right\|^2_{L^2(\Omega)} dt \leq C_J C_u.$$

This proves part a). Next, we see that $\mathcal{T}$ is from $H^1(\Omega)$ and $f := b \cdot \nabla q \in L^2(\Omega_T) + \mathcal{R}[u] \in L^2(\Omega_T)$. After the time transformation $t \mapsto a(T-t)$, we can bring the adjoint equation into the form of $\partial_t \tilde{q} - \Delta \tilde{q} = \frac{1}{a} f$, and hence, the adjoint problem has the form from Theorem 2.3.5 on maximal L^p–regularity of parabolic problems. This yields the higher regularity $q \in L^2(0,T; H^2(\Omega)) \cap H^1(0,T; L^2(0,T))$ which proves the claim b). Next, with this improved regularity, we may deduce essential boundedness with an application of Theorem 2.2.4. For that purpose we observe that $\nabla q \in W(0,T)^d$, and now Corollary 2.1.5 implies that $\nabla q \in L^{2+4/d}(\Omega_T)^d$. □

Under the additional regularity assumptions (F5)–(F7), we can prove (global) Lipschitz continuity for the control–to–adjoint map on the set of admissible controls U^2_{ad}; an analogous assertion can be shown for U^3_{ad}.

Lemma 3.3.3. *(Lipschitz continuity of $\Theta_{|U^2_{\text{ad}}}$ for $d \in \{1,2\}$)*
Let $u_1, u_2 \in U_{\text{ad}} \cap L^2(0,T; H^1(\Omega))$ and let (F1)–(F3), (F5)–(F7) and (J1)–(J2) hold. Let $d \in \{1,2\}$. Then, the following estimate holds

$$\|\Theta(u_1) - \Theta(u_2)\|_{W(0,T)} \leq C_{\text{F}*} C_J C_{\text{ad}} \|u_1 - u_2\|_{L^2(0,T;H^1(\Omega))}.$$

Proof. For $q_1 := \Theta(u_1)$, $q_2 := \Theta(u_2)$, define $\delta u := u_1 - u_2$ and $\delta q := q_1 - q_2$. A quick computation shows that $\delta q \in W(0,T)$ solves the following inhomogeneous parabolic problem

$$\begin{aligned} \langle \partial_t \delta q, \cdot \rangle_{H'} &= \mathcal{F}[u_1](\cdot, \delta q) = F[\delta u] \quad \text{in } L^2(0,T; H^1(\Omega)'), \\ \delta q(T) &= 0 \quad \text{a.e. on } \Omega, \end{aligned}$$

where the r.h.s. is defined f.a.e. $t \in [0,T]$ as

$$F_t[\delta u](\psi) := \langle (\nabla q_2(t)^\top M(t) + \frac{\gamma_1}{2}(u_1 + u_2) + \alpha(t)^\top) \delta u(t), \psi \rangle_{L^2(\Omega)}, \quad \psi \in H^1(\Omega).$$

Hence, in view of Theorem 3.3.2, the following estimate holds

$$\|\delta q\|_{W(0,T)} \leq C_F C_1 \|F[u]\|_{L^2(0,T;H^1(\Omega)')},$$

where $C_1 > 0$ depends continuously only on $\|u_1\|_{L^\infty(\Omega_T)}$. The critical term in $F[\delta u]$ is $\nabla q_2^\top M \, \delta u \, \psi$ since $\nabla q_2(t, \cdot)$ is in general not in $L^\infty(\Omega)$. We partially integrate and use that q_2 fulfills zero Neumann boundary conditions to obtain

$$\int_\Omega \nabla q_2(t,x)^\top M(t,x) \delta u(t,x) \psi(x) \, dx = - \int_\Omega q_2(t,x) \operatorname{div}\big(M(t,x)\, \delta u(t,x)\, \psi(x)\big)\, dx. \tag{3.9}$$

Consequently, we can estimate

$$\begin{aligned} \|F[u]\|_{L^2(0,T;H^1(\Omega)')} \leq & \|\delta u\|_{L^2 H^1} \|q_2\|_{L^\infty(\Omega_T)} \|M\|_{L^\infty W^{1,\infty}} \\ & + \|\delta u\|_{L^2(\Omega_T))} \left(\frac{\gamma_1}{2} \|u_1 + u_2\|_{L^\infty(\Omega_T)} + \|\alpha\|_{L^\infty(\Omega_T)} \right). \end{aligned}$$

Since $u_1, u_2 \in U_{\mathrm{ad}}$, we can estimate $\|u_1 + u_2\|_{L^\infty(\Omega_T)}$ against a constant C_{ad}. Due to the estimate (3.8), we can bound q_2 in the L^∞–norm. The values $\|M\|_{L^\infty W^{1,\infty}}$ and $\|\alpha\|_{L^\infty(\Omega_T)}$ are contained in the constants $C_{\mathrm{F}*}$ and C_J, and obviously $\|\delta u\|_{L^2(\Omega_T))} \leq \|\delta u\|_{L^2 H^1}$. Thus, we obtain

$$\|F[u]\|_{L^2(0,T;H^1(\Omega)')} \leq \|\delta u\|_{L^2 H^1} C_{\mathrm{ad}} C_{\mathrm{F}*},$$

which concludes the proof. □

It can be seen in (3.9) that in the case of only time dependent controls a bound of q_2 in the $L^\infty(0,T;L^2(\Omega))$ norm is sufficient (instead of the $L^\infty(\Omega_T)$ norm.

We continue with our plan to rewrite the terms $\int_{\Omega_T} \mathcal{R}[u]z$, $\int_\Omega \mathcal{T}\, z(T)$ from $\hat{J}'(u)v$ in terms of the adjoint state q. For that purpose, recall that $z := G'(u)v$ for $u, v \in L^\infty(\Omega_T)^m$ solves

$$\begin{aligned} \dot{z} + \mathcal{F}[u](z,\cdot) &= f^{\mathrm{lin}}[u,v], &\qquad& \text{in } L^2(0,T;H^1(\Omega)') \\ z(0) &= 0, && \text{a.e. on } \Omega, \end{aligned}$$

where the r.h.s. of the linearized equation is for $t \in [0,T]$

$$f_t^{\mathrm{lin}}[u,v](\psi) = -\int_\Omega G(u)(t,x)\ (M(t,x)v(t,x)) \cdot \nabla\psi(x)\, dx, \quad \psi \in H^1(\Omega).$$

Comparing this to the weak formulation of the adjoint (3.5), the connection becomes clear. We have

$$\int_{\Omega_T} \mathcal{R}[u](t,x) z(t,x)\, dt\, dx + \int_\Omega \mathcal{T}(x) z(T,x)\, dx = \int_0^T f_t^{\mathrm{lin}}[u,v](q(t))\, dt, \tag{3.10}$$

which follows by testing the weak formulations of z and q, with the $H^1(\Omega)$–functions $q(t)$ and $z(t)$. We remark that, due to the regularity $z, q \in W(0,T)$, it holds that

$$\int_0^T \langle \dot{q}(t), z(t)\rangle_{H'}\, dt = -\int_0^T \langle q(t), \dot{z}(t)\rangle_{H'}\, dt + z(T)q(T).$$

In conclusion, the sought function $\Phi[\bar{u}]$ is obtained from $f^{\mathrm{lin}}[u,v]$ and we have proven the following lemma.

Lemma 3.3.4. *For $u \in L^\infty(\Omega_T)^m$, we define the vector–valued function*

$$\Phi[u] := (\gamma_1 u + \alpha - M\nabla q)\, p \quad \text{on } \Omega_T, \tag{3.11}$$

where $p = G(u)$ is the state and $q = \Theta(u)$ the adjoint. Then, $\Phi \in L^2(\Omega_T)$ and the Fréchet derivative of $\hat{J}$ at u is given by

$$\hat{J}'(u)v = \int_{\Omega_T} \Phi[u] \cdot v\, dt\, dx + \gamma_2 \langle u, v\rangle_Y, \quad v \in \mathcal{U}. \tag{3.12}$$

When we write $\hat{J}'(u) = \Phi[u] + \gamma_2 u$ in the following, we obviously mean it in the sense of (3.12). This fundamental representation allows a detailed first–order analysis.

3.4 Characterization of minimizers – first–order analysis

It is the aim of this section to derive an implicit formula, and prove further regularity properties of local minimizers. Throughout this section, $\bar{u}$ denotes a Y_j–local solution of

$$\min_{u \in U_{\mathrm{ad}}^j} \hat{J}(u).$$

For such $\bar{u}$, the following first–order necessary condition (FONC) holds

$$\int_{\Omega_T} \Phi[\bar{u}] \cdot (v - \bar{u})\, dt\, dx + \gamma_2 \langle \bar{u}, v - \bar{u} \rangle_{U^j} \geq 0, \quad v \in U_{\text{ad}}^j. \tag{3.13}$$

We say that there are inactive constraints if "≥" can be replaced by "=" in (3.13). The following representations of $\bar{u}$ are a direct consequence of this variational inequality; for that purpose recall that

$$Y_1 = L^2(\Omega_T), \quad Y_2 = L^2(0, T; H_0^1(\Omega)), \quad Y_3 = H^1(\Omega_T)$$

and the admissible sets $U_{\text{ad}}^j = U_{\text{ad}} \cap Y_j$ for $j \in \{1, 2, 3\}$.

Corollary 3.4.1. *(The case (U_{ad}^1, Y_1))*
Let $\bar{u} \in U_{\text{ad}}$ be a $L^2(\Omega_T)^m$–local solution.

a) For all $i = 1, \ldots, m$ and a.e. $(t, x) \in \Omega_T$, it holds that

$$\begin{cases} \Phi_i[\bar{u}](t,x) + \gamma_2 \bar{u}_i(t,x) > 0 & \Longrightarrow \; \bar{u}_i(t,x) = u^{\min}, \\ \Phi_i[\bar{u}](t,x) + \gamma_2 \bar{u}_i(t,x) < 0 & \Longrightarrow \; \bar{u}_i(t,x) = u^{\max}, \\ u^{\min} < \bar{u}_i(t,x) < u^{\max} & \Longrightarrow \; \Phi_i[\bar{u}](t,x) + \gamma_2 \bar{u}_i(t,x) = 0. \end{cases}$$

In particular, if $\gamma_2 > 0$ we have the following implicit representation

$$\bar{u}(t,x) = \min\left\{ u^{\max}, \max\left\{ -\frac{1}{\gamma_2} \Phi[\bar{u}](t,x), u^{\min} \right\} \right\}. \tag{3.14}$$

b) In the case of inactive constraints, the following implicit equation holds

$$\gamma_2 \bar{u} = -\left(\gamma_1 \bar{u} + \alpha - M \nabla \bar{q}\right) \bar{p}, \quad \textit{a.e. on } \Omega_T.$$

c) Let $\gamma_2 = 0$ and $\gamma_1 > 0$. In the case of inactive constraints, and if $\bar{p}$ is positive a.e. on Ω_T, we obtain the feedback–like control law

$$\gamma_1 \bar{u} = \alpha - M \nabla \bar{q}, \quad \textit{a.e. on } \Omega_T. \tag{3.15}$$

Proof. Since $\bar{u} \in U_{\text{ad}}^1$ is a U^1–local solution, each component of the FONC reads

$$\langle \Phi_i[\bar{u}] + \gamma_2 \bar{u}_i, v_i - \bar{u}_i \rangle_{L^2(\Omega_T)} \geq 0, \quad v \in U_{\text{ad}}^1.$$

Thus, we can apply Lemma 1.4.6 with $f := \Phi_i[\bar{u}] + \gamma_2 \bar{u}_i \in L^2(\Omega_T)$ and the first claim follows. If $\gamma_2 > 0$, the equation $\Phi[\bar{u}] + \gamma_2 \bar{u} = 0$ can be solved for $\bar{u}$, and we obtain (3.14). Part b) follows trivially by our definition of inactive constraints, that is, $u^{\min} < \bar{u}_i(t,x) < u^{\max}$. Now if $\bar{p}$ is positive, we can divide this implicit equation by $\bar{p}$ and the claim of part c) follows. □

Notice that in part c), $\bar{u}$ is independent of the distribution $\bar{p}$ or initial state p_0 and depends only on functions given by the objective J. This is due to the fact that plugging the formula for $\bar{u}$ into the adjoint problem, we arrive at the following non–linear backward problem for q

$$\begin{aligned} \partial_t q + L^* q &= -\mathcal{R}[\alpha - M \nabla q] && \text{on } \Omega_T, \\ q(T) &= \mathcal{T} && \text{on } \Omega, \\ \nabla q \cdot \hat{n} &= 0 && \text{on }]0, T[\times \partial\Omega, \end{aligned}$$

where the r.h.s. reads

$$\mathcal{R}[\alpha - M\nabla q] = \frac{\gamma_1}{2}|\alpha(t,x) - M\nabla q|^2 + \alpha(t,x) \cdot \big(\alpha(t,x) - M(t,x)\nabla q(t,x)\big) + \beta(t,x).$$

Therefore, solving this non–linear problem for q and calculating the optimal control via (3.15) yields a robust, feedback–like control mechanism in the sense that possible perturbations of $p(t,\cdot)$ or $p_0(\cdot)$ do not result in (possibly) suboptimal controls.

A similar min max–representation for optimal controls in the Y_2 or Y_3 case cannot be expected to hold, since derivatives of the optimal control appear in the FONC. However, if no box constraints are present and $\bar{u} \in L^2(0,T;H_0^1(\Omega))$, boundary conditions on the control are implemented, and in that case, $\bar{u}$ is connected to an elliptic Dirichlet problem.

Corollary 3.4.2. *(The case $L^2(0,T;H_0^1(\Omega))$)*
Let $\gamma_2 > 0$, $d \in \{1,2,3\}$ and let $\bar{u} \in L^2(0,T;H_0^1(\Omega))^m \cap L^\infty(\Omega_T)^m$ be a Y_2–local solution of the minimization problem $\min\{\hat{J}(u) \mid u \in L^2(0,T;H_0^1(\Omega))\}$. Then, $\bar{u}$ has the higher regularity $L^2(0,T;H^2(\Omega))^m$ and solves the elliptic problem

$$-\Delta\bar{u}_i(t,x) + \bar{u}_i(t,x) = -\frac{1}{\gamma_2}\Phi_i[\bar{u}](t,x), \qquad \text{f.a.e. } (t,x) \in \Omega_T, \quad i = 1,\dots,m$$

and f.a.e. $t \in [0,T]$, $u(t,\cdot) = 0$ a.e. on $\partial\Omega$.

We remark that we make no statement on the existence of such minimizer $\bar{u}$ in general.

Proof. Let $i \in \{1,\dots,m\}$. Since $\bar{u}_i \in L^\infty(\Omega_T) \cap L^2(0,T;H_0^1(\Omega))$, it holds that $\bar{p} := G(\bar{u}), \bar{q} := \Theta(\bar{u}) \in W(0,T)$. By an application of Lemma 3.3.4, we can rewrite the the FONC for the Y_2 case as follows

$$\int_{\Omega_T} (\Phi_i[\bar{u}] + \gamma_2\bar{u}_i) \cdot v\,dt\,dx + \gamma_2 \int_{\Omega_T} \nabla\bar{u}_i \cdot v\,dt\,dx = 0, \quad v \in L^2(0,T;H_0^1(\Omega)).$$

Now notice that if $v^1 \in L^2(0,T)$, $v^2 \in H_0^1(\Omega)$, then the product v^1v^2 is in $L^2(0,T;H_0^1(\Omega))$. Therefore, this variational equation can be rewritten as follows

$$\int_0^T \left(\int_\Omega (\Phi_i[\bar{u}] + \gamma_2\bar{u}_i) \cdot v^2\,dx + \gamma_2 \int_\Omega \nabla\bar{u}_i \cdot \nabla v^2\,dx\right) v^1\,dt = 0, \quad v^1 \in L^2(0,T),\ v^2 \in H_0^1(\Omega).$$

Consequently, an application of the fundamental lemma of the calculus of variations on $[0,T]$ implies that f.a.e. $t \in [0,T]$

$$\int_\Omega (\Phi_i[\bar{u}](t) + \gamma_2\bar{u}_i(t)) \cdot v^2\,dx + \gamma_2 \int_\Omega \nabla\bar{u}_i(t) \cdot v^2\,dx = 0, \quad v^2 \in H_0^1(\Omega).$$

Hence, $\bar{u}(t) \in H_0^1(\Omega)$ is a weak solution to the elliptic equation $-\Delta\bar{u}(t,\cdot) + \bar{u}(t,\cdot) = -1/\gamma_2\Phi_i[\bar{u}](t,\cdot)$ on Ω. Let $\bar{p} = G(\bar{u}) \in$ and $\bar{q} = \Theta(\bar{u})$. Since $\bar{p} \in L^\infty(0,T)$, $\nabla\bar{q} \in L^2(\Omega_T)$ and $\Phi[\bar{u}] = (\gamma_1\bar{u} + \alpha - M\nabla q)\,p$ a.e. on Ω_T, it holds that $\Phi_i[\bar{u}](t,\cdot) \in L^2(\Omega)$. An application of the regularity result for elliptic problems, see Lemma 1.4.5, yields the claim. □

In the case of box–constrained controls from $Y_3 = H^1(\Omega_T)$, we cannot expect to deduce an implicit representation for minimizers from the FONC given by (3.13), and hence we perform a second–order analysis for this case in the next section.

3.5 Local uniqueness and coercivity – second–order analysis

We start this section by computing the second–order derivatives of the control–to–state map G from Definition 2.4.1 and the reduced cost functional $\hat{J}$ from (3.1).
Recall from Section 2.4 that $z = G'(u)v$ for $u, v \in L^\infty(\Omega_T)^m$ solves the linearized equation

$$\begin{aligned} \dot{z} + \mathcal{F}[u](z, \cdot) &= f^{\text{lin}}[u, v], && \text{in } L^2(0,T; H^1(\Omega)') \\ z(0) &= 0, && \text{a.e. on } \Omega. \end{aligned}$$

with r.h.s. for $t \in [0,T]$ and $p := G(u)$

$$f_t^{\text{lin}}[u, v](\psi) = -\int_\Omega p(t,x)\ (M(t,x)v(t,x)) \cdot \nabla\psi(x)\, dx, \quad \psi \in H^1(\Omega).$$

We differentiate both equations with respect to u in directions $v_1, v_2 \in L^\infty(\Omega_T)^m$ and obtain the following problem for determining the second–order derivative $w := G''(u)(v_1, v_2) \in W(0,T)$

$$\begin{aligned} \dot{w} + \mathcal{F}[u](w, \cdot) &= f^{\text{quad}}[u, v_1, v_2], && \text{in } L^2(0,T; H^1(\Omega)') \\ w(0) &= 0, && \text{a.e. on } \Omega, \end{aligned}$$

where for $z_i := G'(u)v_i$ for $i = 1, 2$ the right–hand side reads for $t \in [0,T]$ and $\psi \in H^1(\Omega)$

$$f_t^{\text{quad}}[u, v_1, v_2](\psi) := -\int_{\Omega_T} \big(z_1(t,x)M(t,x)v_2(t,x) + z_2(t,x)M(t,x)v_1(t,x)\big) \cdot \nabla\psi(x)\, dx.$$

We remark that $G''(u)(v_1, v_2) = G''(u)(v_2, v_1)$ and if $v = v_1 = v_2$, we simply write $f^{\text{quad}}[u, v]$.
By an application of the chain rule, we can compute the second–order derivative of the reduced cost functional. We have for $u, v_1, v_2 \in U_{\text{ad}}$, for $j = 1, 2, 3$, and $p := G(u), z_1 := G'(u)v_1,\ z_2 := G'(u)v_2$, $w := G''(u)(v_1, v_2)$

$$\begin{aligned} \hat{J}''(u)(v_1, v_2) =& \gamma_1 \int_{\Omega_T} p\, v_1 \cdot v_2\, dt\, dx + \int_{\Omega_T} (\gamma_1 u + \alpha) \cdot (z_2 v_1 + z_1 v_2)\, dt\, dx + \int_{\Omega_T} \mathcal{R}[u] w\, dt\, dx \\ &+ \int_\Omega \mathcal{T} w(T) + \gamma_2 \|v\|_{Y_j}^2. \end{aligned}$$

Analogously to Lemma 3.3.4, we can express $\hat{J}''(\bar{u})(v, v)$ via the adjoint $q := \Theta(u)$ due to

$$\int_{\Omega_T} \mathcal{R}[u] w\, dt\, dx + \int_\Omega \mathcal{T} w(T)\, dt = \int_0^T f_t^{\text{quad}}[u, v](q(t))\, dt = -2\int_{\Omega_T} zMv \cdot \nabla q\, dx\, dt.$$

The main part of this section is to derive the second–order properties of $\hat{J}$ which have been discussed in Section 1.3. More precisely, we want to apply a theorem like Theorem 1.3.5, and for that purpose, let us discuss our possible choices of U_2 and U_∞.
We start the discussion with the case of a L^2 regularizing norm, i.e., the case $Y_1 = L^2(\Omega_T)^m$. We want to point out why it does not seem to be possible to apply Theorem 1.3.5 in that setting. The first and canonical choice is $U_\infty = L^\infty(\Omega_T)^m$ or $U_\infty = U_2 = L^2(\Omega_T)^m$. According to assumption (A1), we need to be able to extend $\hat{J}''(u)$ for some $u \in U_{\text{ad}}$ to a continuous bilinear mapping defined on $U_2 \times U_2$. We will see that in this case, this is in general not possible by considering the term

$$\int_0^T f_t^{\text{quad}}[u, v](q(t))\, dt = -2\int_{\Omega_T} zMv \cdot \nabla q\, dx\, dt.$$

Notice that $z := G'(u)v, q \in W(0,T),\ v \in L^2(\Omega_T)^m$ and $M \in L^\infty(\Omega_T)$. In order for the integral to be finite, z or ∇q need to have higher regularity, but this does, to the best of our knowledge, not hold in general in our setting. We remark that this issue is different in the case of only time–dependent controls.

Since $\nabla q \in L^2(\Omega_T)$, the integral exists if $z \in L^\infty(\Omega_T)$ and Theorem 2.2.3 reveals under which conditions this is the case. The r.h.s. $f^{\text{lin}}[u,v]$ of the equation for z contains the term v which corresponds – more or less – to the function g_2 in Theorem 2.2.3. Therefore, we obtain $z \in L^\infty(\Omega_T)$ if $v \in L^{2+4/d}(\Omega_T)$ where d denotes the dimension of Ω. This is the reason to consider Sobolev spaces as control spaces and use the continuous embedding to higher Lebesgue spaces. Thus, we focus in this section on the third case, where U^3_{ad} is space for the controls and $Y_3 = H^1(\Omega_T)$ is the norm for the regularizing term. Due to the continuous embeddings (2.22), we obtain that for the dimensions $d \in \{1,2\}$, any function $v \in H^1(\Omega_T)$ is also in $L^{2+4/d}(\Omega_T)$. We summarize this and further results in the following lemma.
We repeat that throughout this section, we assume $d = \dim(\Omega) \in \{1,2\}$.

Lemma 3.5.1. *Let $u \in L^\infty \cap H^1(\Omega_T)^m$ and $v_1, v_2 \in H^1(\Omega_T)^m$. Then it holds that*

a) $z_i := G'(u)v_i \in L^\infty(\Omega_T)$ for $i = 1,2$;

b) $w := G''(u)(v_1,v_2) \in L^\infty(\Omega_T)$;

c) the second–order derivative $\hat{J}''(u)(v_1,v_2)$ exists and there exist continuous extensions such that

$$\hat{J}'(u) \in \operatorname{Lin}(H^1(\Omega_T)^m) \quad \textit{and} \quad \hat{J}''(u) \in \operatorname{Bilin}(H^1(\Omega_T)^m \times H^1(\Omega_T)^m).$$

Proof. From the discussion above, it is clear that $v_i \in H^1(\Omega_T)^m$ implies $z_i \in L^\infty(\Omega_T)$. Similarly, we obtain essential bounds for w by considering the r.h.s. of the governing equation

$$f^{\text{quad}}[u,v_1,v_2](\psi) = -\int_\Omega (z_1 M v_2 + z_2 M v_1) \cdot \nabla\psi \, dx \quad \psi \in H^1(\Omega).$$

We want to apply Theorem 2.2.3, and since $v_i \in H^1(\Omega_T)$, $z_i \in L^\infty(\Omega_T)$, $i = 1,2$, we obtain that

$$g_1 = 0, \quad g_2 = z_1 M v_2 + z_2 M v_1 \in L^{2+4/d}(\Omega_T).$$

Since $p, z_i, w \in L^\infty(\Omega_T) \cap W(0,T)$, all the terms of $\hat{J}'(u)v_1$ and $\hat{J}''(u)(v_1,v_2)$ are well–defined and (bi)linear in v_1 and (v_1,v_2), respectively. This concludes the proof. □

Next, we need to verify that condition (C2) holds for $\hat{J}$. For that purpose, we need to establish the Lipschitz continuity of G' and G'' . This is done similarly to Lemma 2.4.3, based on the fact that $z_1 - z_2$ or $w_1 - w_2$ solves again an inhomogeneous Fokker–Planck equation with some right–hand side. Then, according to Corollary 2.2.1 and Theorem 2.2.3, the Lipschitz continuity follows from the convergence to zero of the right–hand side.

Lemma 3.5.2. *(Lipschitz continuity)*
Let $u_1, u_2 \in U_{\text{ad}} \cap H^1(\Omega_T)^m$ and $v \in H^1(\Omega_T)^m$. Define the $W(0,T) \cap L^\infty(\Omega_T)$-functions

$$z_1 := G'(u_1)v, \quad z_2 := G'(u_2)v, \quad w_1 := G''(u_1)(v,v), \quad w_2 := G''(u_2)(v,v).$$

Then, it holds that

$$\begin{aligned}
\|z_1 - z_2\|_{W(0,T)} &\leq C_F C_{\text{ad}} \|v\|_{H^1(\Omega_T)} \|u_1 - u_2\|_{L^2(\Omega_T)}, \\
\|z_1 - z_2\|_{L^\infty(\Omega_T)} &\leq C_F C_{\text{ad}} \|v\|_{H^1(\Omega_T)} \|u_1 - u_2\|_{H^1(\Omega_T)}, \\
\|w_1 - w_2\|_{W(0,T)} &\leq C_F C_{\text{ad}} \|v\|^2_{H^1(\Omega_T)} \|u_1 - u_2\|_{H^1(\Omega_T)}, \\
\|w_1 - w_2\|_{L^\infty(\Omega_T)} &\leq C_F C_{\text{ad}} \|v\|^2_{H^1(\Omega_T)} \|u_1 - u_2\|_{H^1(\Omega_T)},
\end{aligned}$$

where $C_{\text{ad}} > 0$ is a constant that depends continuously only on the box constraints $|u^{\min}|$ and $|u^{\max}|$. Furthermore, $\hat{J}'$ and $\hat{J}''$ are Lipschitz continuous in the sense that

$$\begin{aligned}
|\hat{J}'(u_1)v - \hat{J}'(u_2)v| &\leq C_J C_F C_{\text{ad}} \|v\|_{H^1(\Omega_T)} \|u_1 - u_2\|_{H^1(\Omega_T)}. \\
|\hat{J}''(u_1)(v,v) - \hat{J}''(u_2)(v,v)| &\leq C_J C_F C_{\text{ad}} \|v\|^2_{H^1(\Omega_T)} \|u_1 - u_2\|_{H^1(\Omega_T)}.
\end{aligned}$$

Proof. Define $\delta z := z_1 - z_2 \in W(0,T)$, and observe that δz solves the inhomogeneous Fokker–Planck equation with control u_1 and r.h.s. f_δ^{lin}, that is,

$$\partial_t \delta z + \mathcal{F}[u_1](\delta z, \cdot) = f_\delta^{\text{lin}} \quad \text{in } L^2(0,T;H^1(\Omega'))$$

where we define on $[0,T]$ for $\psi \in H^1(\Omega)$

$$f_\delta^{\text{lin}}(\psi) := f^{\text{lin}}[u_1,v](\psi) - f^{\text{lin}}[u_2,v](\psi) + \int_\Omega z_2\, M(u_1-u_2)\cdot\nabla\psi\, dx.$$

According to Corollary 2.2.1, we need to bound f_δ^{lin} in $L^2(0,T;H^1(\Omega)')$ for the $W(0,T)$–estimate. Almost everywhere on $[0,T]$, we obtain the following estimate, where $p_1 := G(u_1)$, $p_2 := G(u_2)$.

$$\begin{aligned}|f_\delta^{\text{lin}}(\psi)| &\le \int_\Omega |(p_1 M v - p_2 M v)\cdot\nabla\psi|\; dx + \int_\Omega |z_2 M(u_1-u_2)\cdot\nabla\psi|\; dx\\ &\le \|M\|_\infty \|\nabla\psi\|_2 \Big(\|p_1-p_2\|_\infty \|v\|_2 + \|z_2\|_\infty \|u_1-u_2\|_2\Big).\end{aligned}$$

Since we have the estimates $\|p_1 - p_2\|_{L^\infty(\Omega_T)} \le C_F C_1 C_2 \|u_1-u_2\|_{L^2(\Omega_T)}$ by Lemma 2.4.3 d) and $\|z_2\|_{L^\infty(\Omega_T)} \le C_F C_2 \|v\|_{H^1(\Omega_T)}$ by Lemma 3.5.1, where $C_i > 0$ depends continuously only on $\|u_i\|_{L^\infty(\Omega_T)}$ for $i \in \{1,2\}$, we deduce that

$$\|f_\delta^{\text{lin}}(\psi)\|_{L^2(0,T;H^1(\Omega)')} \le C_F C_1 C_2 \|v\|_{H^1(\Omega_T)} \|u_1-u_2\|_{L^2(\Omega_T)}.$$

This gives, according to Corollary 2.2.1, the Lipschitz bound in the $W(0,T)$–norm. The L^∞–estimate is obtained with Theorem 2.2.3 if we bound the term $g_2 := (p_1-p_2)Mv + z_2 M(u_1-u_2)$ of the r.h.s, i.e. $f_\delta^{\text{lin}}(\psi) = \int_\Omega g_2\cdot\nabla\psi$, in the $L^q(\Omega_T)$–norm for $q = d+2$ as follows

$$\begin{aligned}\|g_2\|_{L^q(\Omega_T)} &= \|(p_1-p_2)Mv + z_2\, M(u_1-u_2)\|_{L^q(\Omega_T)}\\ &\le C_F\Big(\|p_1-p_2\|_{L^\infty(\Omega_T)}\|v\|_{L^q(\Omega_T)} + \|z_2\|_{L^\infty(\Omega_T)}\|u_1-u_2\|_{L^q(\Omega_T)}\Big).\end{aligned}$$

The claim follows from the continuous embedding from $H^1(\Omega_T)$ into $L^q(\Omega_T)$.
Once we have established the Lipschitz continuity for G' in the $W(0,T)$ and $L^\infty(\Omega_T)$–norm, the same procedure can be done with $\delta w := w_1 - w_2 \in W(0,T)$ in order to obtain the Lipschitz estimates for G''. For any $\psi \in H^1(\Omega)$ and a.e. on $[0,T]$, we find that

$$\begin{aligned}\langle\partial_t(\delta w),\psi\rangle_{H'} + &\mathcal{F}[u_1](\delta w,\psi)\\ &= f^{\text{quad}}[u_1,v_1](\psi) - f^{\text{quad}}[u_2,v_2](\psi) + \int_\Omega w_2 M(u_1-u_2)\cdot\nabla\psi\, dx =: f_\delta^{\text{quad}}(\psi),\end{aligned}$$

where $f_\delta^{\text{quad}} \in L^2(0,T;H^1(\Omega)')$ can be bound analogously to f_δ^{lin} from above. Thus, we have shown the first claim. This immediately yields the Lipschitz continuity of $\hat{J}'$ and $\hat{J}''$; we will only prove it for the latter case. We have

$$\begin{aligned}|\hat{J}''(u_1)(v,v) - \hat{J}''(u_2)(v,v)| = \Big|&\gamma_1\int_{\Omega_T}(p_1-p_2)|v|^2 dt\, dx + 2\int_{\Omega_T}((\gamma_1 u_1+\alpha)z_1 - (\gamma_1 u_2+\alpha)z_2)\cdot v\, dt\, dx\\ &+ \int_{\Omega_T}(\mathcal{R}[u_1]w_1 - \mathcal{R}[u_2]w_2)\, dt\, dx + \int_\Omega(\mathcal{T}w_1(T) - \mathcal{T}w_2(T))\Big|\\ \le&\gamma_1\|p_1-p_2\|_{L^\infty(\Omega_T)}\|v\|_{L^2(\Omega_T)}\\ &+ 2\Big(\|\alpha\|_{L^\infty(\Omega_T)}\|z_1-z_2\|_{L^\infty(\Omega_T)} + \gamma_1\|u_1z_1 - u_2z_2\|_{L^2(\Omega_T)}\Big)\|v\|_{L^2(\Omega_T)}\\ &+ \|\mathcal{R}[u_1]\|_{L^1(\Omega_T)}\|w_1-w_2\|_{L^\infty(\Omega_T)} + \|w_2\|_{L^\infty(\Omega_T)}\|\mathcal{R}[u_1]-\mathcal{R}[u_2]\|_{L^1(\Omega_T)}\\ &+ \|\mathcal{T}\|_{L^2(\Omega)}\|w_1(T)-w_2(T)\|_{L^2(\Omega)}.\end{aligned}$$

In order to treat the bilinear term, we just observe that

$$\|u_1 z_1 - u_2 z_2\|_{L^2(\Omega_T)} \leq \|u_1\|_{L^2(\Omega_T)} \|z_1 - z_2\|_{L^\infty(\Omega_T)} + \|z_2\|_{L^\infty(\Omega_T)} \|u_1 - u_2\|_{L^2(\Omega_T)}$$

and estimate $\|z_1 - z_2\|_{L^\infty(\Omega_T)} \leq C\|u_1 - u_2\|_{H^1(\Omega_T)}$. This concludes the proof. □

Next, we can verify condition (C2) in order to apply Theorem 1.3.6. The choice for the spaces are

$$U_2 = U_\infty = H^1(\Omega_T)^m, \quad A = L^\infty \cap H^1(\Omega_T)^m,$$

and U_{ad}^3 is the set of admissible controls.

Lemma 3.5.3. *The reduced objective $\hat{J}$ fulfills condition (C2).*

Proof. First, we notice that due to the Lipschitz continuity it is enough to verify (C2) for a fixed control u instead of a sequence (u^k) as in Theorem 1.3.5. Thus, let $u \in L^\infty \cap H^1(\Omega_T)^m$ and $(v^k)_{k\in\mathbb{N}} \subset H^1(\Omega_T)^m$ with $v^k \rightharpoonup v$ weakly in $H^1(\Omega_T)^m$.
Condition (C2.1), that is

$$\hat{J}'(u)v^k = \int_{\Omega_T} \Phi[u] \cdot v^k \, dt \, dx + \gamma_2 \langle u, v^k \rangle_{H^1(\Omega_T)} \to \hat{J}'(u)v, \quad k \to \infty,$$

follows immediately from the representation given in Lemma 3.12, since $\Phi[u] \in L^2(\Omega_T)^m$, $u \in H^1(\Omega_T)^m$ and due to the weak convergence of (v^k).
Next let $z_k := G'(u)v^k$ and consider the second–order derivative of $\hat{J}$

$$\hat{J}''(u)(v^k, v^k) = \gamma_1 \int_{\Omega_T} p\,|v^k|^2 \, dt \, dx + 2\int_{\Omega_T} (\gamma_1 u + \alpha) z_k \cdot v^k \, dt \, dx - 2\int_{\Omega_T} z_k (Mv^k) \cdot \nabla q \, dx \, dt + \gamma_2 \|v^k\|_{Y_3}^2.$$

For the first term, the weak convergence $v^k \rightharpoonup v$ in $L^2(\Omega_T)^m$ is sufficient. This can be seen by an application of Lemma 1.4.1 with $p \in L^\infty(\Omega_T)$; we therefore obtain

$$\gamma_1 \int_{\Omega_T} p\,|v|^2 \, dt \, dx \leq \liminf_{k\to\infty} \gamma_1 \int_{\Omega_T} p\,|v^k|^2 \, dt \, dx. \tag{3.16}$$

In view of the second–term, notice that $z_k \to z = G'(u)v$ strongly in $L^2(\Omega_T)$ and $(\gamma_1 u + \alpha) \in L^\infty(\Omega_T)^m$. Hence, the weak convergence of $v^k \rightharpoonup v$ in $L^2(\Omega_T)$ is again sufficient to deduce

$$\lim_{k\to\infty} \int_{\Omega_T} (\gamma_1 u + \alpha) z_k \cdot v^k \, dt \, dx = \int_{\Omega_T} (\gamma_1 u + \alpha) z \cdot v \, dt \, dx.$$

The third term is obviously the critical one. Since $(\nabla q)^\top M \in L^2(\Omega_T)^m$ and $v^k \rightharpoonup v$ in $H^1(\Omega_T)^m$, we need at least that $z_k \to z$ in $L^r(\Omega_T)$, where the exponent r satisfies $1/r + 1/q + 1/2 = 1$ and q satisfies $H^1(\Omega_T) \subset L^q(\Omega_T)$. Since we assume the dimension of Ω to be in $\{1, 2\}$, we can simply choose $r = \infty$ and apply Lemma 2.4.2. This yields

$$\lim_{k\to\infty} \int_{\Omega_T} z_k \nabla q^\top M v^k \, dx \, dt = \int_{\Omega_T} z \nabla q^\top M v \, dx \, dt.$$

Lastly for the fourth term, we exploit the weak lower semicontinuity to obtain

$$\gamma_2 \|v^k\|_{H^1(\Omega_T)}^2 \leq \liminf_{k\to\infty} \gamma_2 \|v\|_{H^1(\Omega_T)}^2.$$

Therefore, we have verified the conditions (C1)–(C2.2) for our optimal control problem in the $H^1(\Omega_T)^m$–setting. In order for condition (C2.3) to be satisfied, it appears that γ_2 must be positive (but may be arbitrarily small). In that case, we have $v^k \rightharpoonup 0$ in $H^1(\Omega_T)^m$ and therefore $z_k \to 0$ in $L^\infty(\Omega_T)$. Due to the Lipschitz continuity of $\hat{J}''$, we are again allowed to consider the stationary sequence u for the controls in (C2.3). Applying the estimate (3.16), we obtain for $\Lambda := \gamma_2$

$$\Lambda \liminf_{k\to\infty} \|v^k\|^2_{H^1(\Omega_T)} \leq \liminf_{k\to\infty} \hat{J}''(u)(v^k, v^k).$$

This concludes the proof. □

Now we can state the main result on the second–order analysis of the optimal control problem. For that purpose, let $\bar{u}$ be a local minimizer of

$$\min_{u \in U^3_{\text{ad}}} \hat{J}(u), \tag{3.17}$$

and recall for $\tau > 0$ the sets

$$\begin{aligned} S_{\bar{u}} &= \big\{\lambda(u - \bar{u}) \,:\, \lambda > 0 \text{ and } u \in U^3_{\text{ad}}\big\}, && \text{(cone of feasible directions)} \\ C_{\bar{u}} &= \overline{S_{\bar{u}}}^{H^1(\Omega_T)^m} \cap \big\{v \in H^1(\Omega_T)^m \,:\, \hat{J}'(\bar{u})v = 0\big\}, && \text{(critical cone)} \\ E^\tau_{\bar{u}} &= \left\{ v \in \overline{S_{\bar{u}}}^{H^1(\Omega_T)^m} \,:\, |\hat{J}'(\bar{u})v| \leq \tau \|v\|_2 \right\} && \text{(extended cone).} \end{aligned}$$

Furthermore, we repeat that the assumptions (J1)–(J2) on J, defined in (3.1), are given in the beginning of this chapter. The assumptions of (F1)–(F3) on the Fokker–Planck problem, see Definition 2.1.1, is formulated at the beginning of Chapter 2. The sets of admissible controls under consideration are defined in (2.35) and in the beginning of Section 2.4.

Theorem 3.5.4. *(Main result on second–order sufficient conditions)*
Let (F1)–(F3) and (J1)–(J2) hold. Let $\bar{u}$ satisfy (A1) *and* (A2) *from Theorem 1.3.6. Then, there exists $\varepsilon, \delta, \nu, \tau > 0$ such that the following holds.*

a) For all $u \in U^3_{\text{ad}} \cap B_\varepsilon(\bar{u}; H^1(\Omega_T)^m)$, it holds that

$$\hat{J}(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|^2_{H^1(\Omega_T)} \leq \hat{J}(u).$$

b) For all critical points $u^* \in U_{\text{ad}} \cap B_\varepsilon(\bar{u}; H^1(\Omega_T)^m)$, it holds that

$$\bar{u} = u^*.$$

c) For all $u \in U_{\text{ad}} \cap B_\varepsilon(\bar{u}; H^1(\Omega_T)^m)$ and all $v \in E^\tau_{\bar{u}}$, it holds that

$$\hat{J}''(u)(v, v) \geq \frac{\nu}{2}\|v\|^2_{H^1(\Omega_T)}.$$

Proof. We only have to combine our previous results in order to prove the claim: Let us consider Theorem 1.3.6 and choose

$$U_2 = U_\infty = H^1(\Omega_T)^m, \quad A = L^\infty \cap H^1(\Omega_T)^m.$$

Hence, (C1) and (C2) are fulfilled according to Lemma 3.5.1 and 3.5.3. Consequently, we may apply Theorem 1.3.6 which concludes the proof. □

4

Fokker–Planck optimal control problems of tracking type

Classification of mathematical problems as linear and non–linear is like classification of the Universe as bananas and non–bananas.

Source unknown

This chapter is devoted to a first– and second–order analysis of a Fokker–Planck optimal control problem

$$\min_{u\in\mathcal{U}} J(p,u) \tag{4.1}$$

$$\partial_t p + \mathcal{F}[u](p,\cdot) = 0 \quad \text{in } L^2(0,T;H^1(\Omega)'), \tag{4.2}$$

$$p(0) = p_0 \quad \text{in } L^2(\Omega), \tag{4.3}$$

where the cost functional is of tracking type

$$J(p,u) = \frac{\beta}{2}\|p - p^d\|^2_{L^2(\Omega_T)} + \frac{\alpha}{2}\|p(T) - p^T\|^2_{L^2(\Omega)} + \frac{\gamma}{2}\|u\|^2_Y. \tag{4.4}$$

We assume the diffusion $a > 0$ to be constant, and we recall the bilinear flux–operator

$$\mathcal{F}_t : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R} \quad \text{f.a.e. } t \in]0,T[\,,$$
$$\mathcal{F}_t(p,\psi) := \int_\Omega \big(a\nabla p(x)\cdot\nabla\psi(x) - p(x)\,B[u](t,x)\cdot\nabla\psi(x)\big)\,dx.$$

We focus only on time–dependent controls from the spaces $L^2(0,T)^m$ and from $H^1_0(0,T)^m$. For this chapter, we consider the following two box–constrained sets

$$U^T_{\mathrm{ad}} \quad \text{and} \quad U^{T,H}_{\mathrm{ad}} := U^T_{\mathrm{ad}} \cap H^1_0(0,T)^m.$$

Furthermore, in this setting, the optimal control problem (4.1) is also well–posed in the absence of box constraints, and therefore, the possible choices of the set of admissible controls $\mathcal{U}$, and regularizing norms are

$$\mathcal{U} \in \left\{ H_0^1(0,T)^m, U_{\text{ad}}^T, U_{\text{ad}}^{T,H} \right\} \quad \text{and} \quad \|\cdot\|_Y = \|\cdot\|_2 \text{ or } \|\cdot\|_{H^1}.$$

In a tracking type formulation of the objective, the aim is to drive the PDF p to a desired distribution p^d defined on Ω_T with terminal distribution p^T defined on Ω. Another application of problem 4.1–4.4 stems from the field of inverse problems. In this setting, it is the aim to find (or identify) the drift (or parts of the drift) knowing p^d and p^T from measurements. The weights α, β are assumed to be non–negative, and the regularizing term γ is assumed to be positive. We impose the following regularity conditions on J

$$p^d \in H^2(\Omega_T), \quad p^T \in H^2(\Omega). \tag{J1}$$

For most results, L^2–regularity of p^d and p^T would be sufficient, however, for the second–order analysis and for sufficient accuracy rates of the Galerkin discretization presented in the following chapters, the higher regularity (J1) is necessary. We denote with $C_J > 0$ a constant that depends continuously on $\alpha, \beta, \|p^d\|_{H^2(Q)}$ and $\|p^T\|_{H^2(\Omega)}$. We remark that C_F and C_F* are constants that depend continuously on the quantities given in (F1)–(F3) and (F1)–(F7), respectively, and refer the reader to Chapter 2. We recall the box–constraint constant $C_{\text{ad}} > 0$ depending continuously only on $|u^{\min}|, |u^{\max}|$. It is the aim of this chapter to show the existence of optimal controls, derive the optimality system and analyze the adjoint problem. Furthermore, we derive an implicit representation of local minimizer and show higher regularity of optimal controls and the corresponding states. Then, we show second–order results by an application of Theorem 1.3.6.

We remark that an analogous first– and second–order analysis of the control problem with a tracking type objective can be performed for time–space dependent controls with the techniques from the previous chapter. The main difficulties are more or less the same, and most proofs can be done analogously. However, we prefer to keep this chapter clear and compact, and thus, we focus only on time–dependent controls. Further, we have chosen to split this chapter in a section of results and a section of proofs for a better overview and due to the fact that some results have already been established in [5].

4.1 Main results

Throughout this chapter, the assumptions (F1)–(F7) hold for the FP problem, given in the beginning of Chapter 2. The spatial domain Ω is convex and is polygonal or has sufficiently smooth boundary and may have arbitrary dimension $d \in \mathbb{N}$. We assume that M, c are only space–dependent, hence the drift is of the form

$$B[u](t,x) = M(x)u(t) + c(x), \quad (t,x) \in \Omega_T.$$

This ansatz can be seen as an m–dimensional approximation to a time–space dependent control mechanism, where the space dependency is chosen a–priori with M and c, and the m–dimensional control u is a regulating function. The classical Fokker–Planck equation for p with a constant diffusion $a > 0$ reads

$$\partial_t p - a\Delta p + \operatorname{div}\big(B[u]p\big) = 0 \qquad \text{on } \left]0,T\right[\times \Omega$$

with flux–zero boundary conditions

$$(a\nabla p - B[u]p) \cdot \hat{n} = 0 \qquad \text{on } \left]0,T\right[\times \partial\Omega.$$

In Section 2.1 and 2.3, we have shown that the control–to–state map G, under the assumptions from above, satisfies

$$G : L^2(0,T)^m \to W(0,T) \quad \text{and} \quad G : L^\infty(0,T)^m \to H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H^2(\Omega)).$$

Due to Lemma 2.3.2, it holds that $G : L^\infty(0,T)^m \to C([0,T]; L^2(\Omega))$ is compact. The Fréchet differentiability of G on $L^\infty(0,T)^m$ has been established in Section 2.4. Recall that $z := G'(u)v$ and $w := G''(u)(v,v)$ are solutions to

$$\begin{aligned} &\langle \dot{z}, \cdot \rangle_{H'} + \mathcal{F}[u](z,\cdot) = \langle f^{\mathrm{lin}}[u,v], \cdot \rangle_{H'} && \text{in } L^2(0,T;H^1(\Omega)'), \\ &z(0) = 0 && \text{a.e. on } \Omega. \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} &\langle \dot{w}, \cdot \rangle_{H'} + \mathcal{F}[u](w,\cdot) = \langle f^{\mathrm{quad}}[u,v], \cdot \rangle_{H'} && \text{in } L^2(0,T;H^1(\Omega)'), \\ &w(0) = 0 && \text{a.e. on } \Omega, \end{aligned} \tag{4.6}$$

where for $\psi \in H^1(\Omega)$, $t \in [0,T]$, the right–hand sides read

$$\langle f_t^{\mathrm{lin}}[u,v], \psi \rangle_{H'} = -\int_\Omega p(t,x)\, v(t)^\top M(x) \nabla \psi(x)\, dx. \tag{4.7}$$

$$\langle f_t^{\mathrm{quad}}[u,v], \psi \rangle_{H'} = -2\int_\Omega z(t,x)\, v(t)^\top M(x) \nabla \psi(x)\, dx. \tag{4.8}$$

Since controls are now only time–dependent and M, p, z fulfill the Neumann boundary condition on $\partial\Omega$, we can partially integrate (4.7) and (4.8). Thus, we deduce that the right–hand sides can be represented by L^2–functions which are denoted in the same way

$$f^{\mathrm{lin}}[u,v], f^{\mathrm{quad}}[u,v] \in L^2(\Omega_T).$$

Consequently, we obtain improved regularity of the Fréchet derivatives, and the following result will be useful for the analysis of the Galerkin discretization of the optimal control problem.

Theorem 4.1.1. *(Improved regularity of z and w)*
Let $u \in U_{\mathrm{ad}}^T$, $v \in L^2(0,T)^m$, $p := G(u)$. Then, $z := G'(u)v$ and $w := G''(u)(v,v)$ satisfy the estimate

$$\|z\|_{L^2H^2} + \|w\|_{L^2H^2} + \|z\|_{H^1L^2} + \|w\|_{H^1L^2} \le C_{\mathrm{ad}} C_{\mathrm{F}*} C_v, \tag{4.9}$$

where $C_v > 0$ depends continuously only on $\|v\|_{L^2(0,T)}$.

Next, we investigate the existence of optimal controls for $\hat{J}(\cdot) := J(G(\cdot), \cdot)$.

Theorem 4.1.2. *The optimal control problem*

$$\min_{u \in \mathcal{U}} \hat{J}(u)$$

possesses a global minimizer $u^* \in \mathcal{U}$ for $\mathcal{U} \in \left\{ H_0^1(0,T)^m, U_{\mathrm{ad}}^T, U_{\mathrm{ad}}^{T,H} \right\}$.

The Fréchet differentiability of G from $L^\infty(0,T)$ to $W(0,T)$ implies that $\hat{J}$ is arbitrarily often Fréchet differentiable on $L^\infty(0,T)^m$. Since $H_0^1(0,T)^m \subset L^\infty(0,T)^m$, we obtain the following result.

Theorem 4.1.3. *The reduced objective $\hat{J}$ is of class C^2 on $L^\infty(0,T)^m$ and $H_0^1(0,T)^m$ with derivatives*

$$\begin{aligned} \hat{J}'(u)v &= \beta \int_{\Omega_T} (p - p^d) z \, dx\, dt + \alpha \int_\Omega (p(T) - p^T)\, dx + \gamma \langle u, v \rangle_Y, \\ \hat{J}''(u)(v,v) &= \beta \|z\|^2_{L^2(\Omega_T)} + \beta \int_{\Omega_T} (p - p^d) w \, dx\, dt \\ &\quad + \alpha \|z(T)\|^2_{L^2(\Omega)} + \alpha \int_\Omega (p(T) - p^T) w(T)\, dx + \gamma \|v\|_Y^2, \end{aligned}$$

where $p = G(u)$, $z = G'(u)v$, $w = G''(u)(v,v)$ and $Y = L^2(0,T)^m$ or $H^1(0,T)^m$, respectively.

We can represent the reduced gradient $v \mapsto \hat{J}'(u)v$ at u with the adjoint function q at u given by the following backward problem

Definition 4.1.4. *The function $q \in W(0,T)$ is a solution to the adjoint problem with control $u \in L^\infty(0,T)^m$ and state $p = G(u)$ if*

$$\begin{aligned} -\partial_t q + \mathcal{F}[u](\cdot, q) &= \beta(p - p^d) && \text{in } L^2(0,T;H^1(\Omega)') \\ q(T) &= \alpha(p(T) - p^T) && \text{a.e. on } \Omega. \end{aligned}$$

In that case, q is called the adjoint state associated with (p,u). We notice that there are different sign conventions for q.

The control–to–adjoint map $\Theta : L^\infty(0,T)^m \to W(0,T)$ is well–posed and Lipschitz continuous in $W(0,T)$, uniformly on U_{ad}^T. Furthermore, we have the higher regularity

$$\Theta : L^\infty(0,T)^m \to H^1(0,T;L^2(\Omega)) \cap C([0,T];H^1(\Omega)) \tag{4.10}$$

and the compactness of $\Theta : L^\infty(0,T)^m \to C([0,T];L^2(\Omega))$.
With the adjoint, we can rewrite the reduced gradient, using the fact that

$$\int_{\Omega_T} p\nabla q^\top M v \, dx\, dt = -\beta \int_{\Omega_T} (p - p^d) z \, dx\, dt - \alpha \int_\Omega (p(T) - p^T) z(T) \, dx\, dt,$$

as follows

$$\hat{J}'(u)v = -\int_{\Omega_T} p\nabla q^\top M v \, dx\, dt + \gamma \langle u, v\rangle_Y.$$

Analogously to Chapter 3, for $u \in L^\infty(0,T)^m$, we introduce the function $\Phi[u] : [0,T] \to \mathbb{R}^m$,

$$\Phi[u](t) := -\int_\Omega p(t,x)\nabla q(t,x)^\top M(x) \, dx, \tag{4.11}$$

where $p = G(u)$ and $q = \Theta(u)$. Due to the regularity $p \in C([0,T];H^1(0,T))$ and $\nabla q \in C([0,T];L^2(\Omega))^d$, it holds that (after modification on a set of measure zero)

$$\Phi[u] \in C([0,T])^m.$$

Furthermore, for the case $Y = L^2(0,T)^m$ and $u \in L^\infty(0,T)^m$, we obtain the pointwise representation of the Fréchet derivative

$$\hat{J}'(u)(t) = \Phi[u](t) + \gamma u(t), \quad \text{f.a.e. } t \in [0,T].$$

The optimality system for the triplet (u,p,q) in a classical formulation reads

$$\begin{aligned} \partial_t p - a\Delta p + \operatorname{div}\big(B[u]p\big) &= 0 && \text{on } [0,T] \times \Omega, \\ p(0) &= p_0 && \text{on } \Omega, \\ (a\nabla p - B[u]p) \cdot \hat{n} &= 0 && \text{on } [0,T] \times \partial\Omega, \\ -\partial_t q - a\Delta q + B[u] \cdot \nabla q &= \beta(p - p^d) && \text{on } [0,T] \times \Omega, \\ q(T) &= \alpha(p(T) - p^T) && \text{on } \Omega, \\ \nabla q \cdot \hat{n} &= 0 && \text{on } [0,T] \times \partial\Omega, \end{aligned}$$

and for all $v \in \mathcal{U}$ it holds that

$$\begin{cases} \hat{J}'(u)(v - u) \geq 0, & \text{if } \mathcal{U} = U_{\text{ad}}^T \text{ or } U_{\text{ad}}^{T,H}, \\ \hat{J}'(u)v = 0, & \text{if } \mathcal{U} = H_0^1(0,T)^m. \end{cases}$$

Next, we can state the implicit representation for a local minimizer

Theorem 4.1.5. *(Implicit formula for local minimizer and higher regularity)*

a) Let $\bar{u}$ be a local minimizer of the problem

$$\min_{u \in U_{\mathrm{ad}}^T} \hat{J}(u)$$

with $\|\cdot\|_Y = \|\cdot\|_2$. Then it holds for $i = 1, \dots, m$

$$\begin{cases} \Phi_i[\bar{u}](t) + \gamma \bar{u}_i(t) > 0 & \Longrightarrow \bar{u}_i(t) = u^{\min}, \\ \Phi_i[\bar{u}](t) + \gamma \bar{u}_i(t) < 0 & \Longrightarrow \bar{u}_i(t) = u^{\max}, \\ u^{\min} < \bar{u}_i(t) < u^{\max} & \Longrightarrow \Phi_i[\bar{u}](t) + \gamma \bar{u}_i(t) = 0. \end{cases}$$

and $\bar{u}$ is continuous with the following representation

$$\bar{u}(t) = \min\left\{ u^{\max}, \max\left\{ -\frac{1}{\gamma}\Phi[\bar{u}](t), u^{\min} \right\}\right\}. \tag{4.12}$$

Furthermore, in the case of inactive constraints, this becomes

$$\bar{u}(t) = \frac{1}{\gamma}\int_\Omega \bar{p}(t,x)\nabla\bar{q}(t,x)^\top M(x)\,dx, \quad t \in [0,T].$$

b) Let $\bar{u}$ be a local minimizer of the problem

$$\min_{u \in H_0^1(0,T)^m} \hat{J}(u)$$

with $\|\cdot\|_Y = \|\cdot\|_{H^1}$. Then $\bar{u}$ has the higher regularity $C^2(0,T)^m \cap H_0^1(0,T)^m$ and satisfies for $t \in [0,T]$ the boundary value problem (or elliptic Dirichlet problem)

$$\bar{u}''(t) = \bar{u}(t) + \frac{1}{\gamma}\Phi[\bar{u}](t), \quad \bar{u}(0) = 0 = \bar{u}(T).$$

Next, we analyze the second–order conditions for the case U_{ad}^T, which was done in [5].

Theorem 4.1.6. *The reduced objective $\hat{J}$ satisfies the conditions (C1)–(C2) of Theorem 1.3.6 for*

$$\text{the admissible set } U_{\mathrm{ad}}^T \quad \text{and} \quad U_2 = U_\infty = A = L^2(0,T)^m.$$

Let $\bar{u}$ satisfy (A1)–(A2). Hence, there exists $\varepsilon, \delta, \nu, \tau > 0$ such that

a) for all $u \in U_{\mathrm{ad}}^T \cap B_\varepsilon(\bar{u}; L^2(0,T))$ it holds that

$$\hat{J}(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|_2^2 \le J(u),$$

b) for all critical points u^* with $u^* \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; L^2(0,T))$ it holds that

$$\bar{u} = u^*,$$

c) for all $u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; L^2(0,T))$ and all $v \in E_{\bar{u}}^\tau$ it holds that

$$\hat{J}''(u)(v,v) \ge \frac{\nu}{2}\|v\|_2^2,$$

where $E_{\bar{u}}^\tau = \left\{ v \in \overline{S_{\bar{u}}}^{L^2(0,T)} : |J'(\bar{u})v| \le \tau\|v\|_2 \right\}$.

4.2 Proofs

We start with the higher regularity of z and w.

Proof of Theorem 4.1.1. Recall that (F1)–(F7) hold, and therefore, $f^{\text{lin}}[u,v]$ and $f^{\text{quad}}[u,v]$ can be represented as $L^2(\Omega_T)$–functions. Thus, we can bring the equations for z and w into the form of Theorem 2.3.5 on maximal L^p–regularity of parabolic problems. After estimating $\|f^{\text{lin}}[u,v]\|_2 \leq C_{\text{ad}} C_{\text{F}*} \|v\|_2$ and $\|f^{\text{quad}}[u,v]\|_2 \leq C_{\text{ad}} C_{\text{F}*} \|v\|_2^2$, an application of this theorem yields the claim. □

Next, we prove the existence of optimal controls.

Proof of Theorem 4.1.2. Let us begin with the box–constrained cases $\mathcal{U} = U_{\text{ad}}^T$ or $U_{\text{ad}}^{T,H}$. Since $\hat{J}$ is non–negative, it is bounded from below by zero, and we can pick a minimizing sequence $(u^k)_{k\in\mathbb{N}}$ from the set of admissible controls $\mathcal{U}$ with

$$\hat{J}(u^k) \to \mathcal{I} := \inf_{u\in\mathcal{U}} \hat{J}(u), \quad \text{as } k \to \infty.$$

Due to the box–constraints, (u^k) is bounded uniformly in $L^\infty(0,T)$ and has a weak*–limit $\bar{u} \in \mathcal{U}$. Therefore, by the compactness of $G : L^\infty(0,T)^m \to C([0,T]; L^2(\Omega))$, we have for a subsequence

$$G(u^k) \to G(u) \text{ in } L^\infty(0,T; L^2(\Omega)),$$

which immediately implies

$$\|G(u^k) - p^d\|_{L^2(\Omega_T)}^2 \to \|G(u) - p^d\|_{L^2(\Omega_T)}^2 \quad \text{and} \quad \|G(u^k)(T) - p^T\|_{L^2(\Omega)}^2 \to \|G(u) - p^T\|_{L^2(\Omega)}^2.$$

If the regularizing norm Y is H^1, then we additionally make use of the lower boundedness of $\big(\hat{J}(u^k)\big)_{k\in\mathbb{N}}$ which yields the weak convergence of $(u^k)_{k\in\mathbb{N}}$ in H^1 in that case (after possibly extracting a subsequence). All in all, this implies the weak lower semicontinuity of $\hat{J}$, and therefore, it holds that

$$\mathcal{I} \leq \hat{J}(\bar{u}) \leq \liminf_{k\to\infty} \hat{J}(u^k) = \mathcal{I}.$$

Thus, $\bar{u}$ is a minimizer of $\hat{J}$.
The unconstrained case, i.e., the set of admissible controls is $H_0^1(0,T)^m$, follows completely analogously, except the proof of the boundedness of the minimizing sequence. In this case, we obtain it due to the positivity of γ and the non–negativity of $\hat{J}$, as follows. Without loss of generality, we may assume that the uncontrolled case $u = 0$ is not optimal, i.e. $\hat{J}(0) > \mathcal{I}$. Thus, for sufficiently large k, it holds that $\hat{J}(0) \geq \hat{J}(u^k)$, which implies that $\frac{\gamma}{2}\|u^k\|_{H^1}^2$ is uniformly bounded by a function of the uncontrolled state $G(0)$. Then, due to the continuous embedding $H_0^1(0,T)^m \hookrightarrow L^\infty(0,T)^m$, we can estimate

$$\frac{\gamma}{2}\|u^k\|_\infty^2 \leq C_T \left(\frac{\beta}{2}\|G(0) - p^d\|_{L^2(\Omega_T)}^2 + \frac{\alpha}{2}\|G(0)(T) - p^T\|_{L^2(\Omega)}^2 \right) \leq (\alpha+\beta) C_F C_J,$$

where the constant $C_F > 0$ was introduced at the beginning of Chapter 2. This concludes the proof. □

Next, we need to verify the higher regularity of the adjoint stated in (4.10).

Proof of (4.10). After time transformation $t \mapsto T - t$, that is $q := \Theta(u)(T-t)$, we see that the adjoint problem can be transformed in a forward heat problem with Neumann–boundary conditions

$$\begin{aligned} \partial_t q - a\Delta q &= f && \text{on } [0,T] \times \Omega, \\ q(0) &= q_0 && \text{on } \Omega, \\ \nabla q \cdot \hat{n} &= 0 && \text{on } [0,T] \times \partial\Omega, \end{aligned}$$

with the following initial data and right–hand side

$$q_0 := \alpha(p(T) - p^T), \qquad f := \beta(p - p^d) - B[u] \cdot \nabla q.$$

Since we have already established higher regularity for the Fokker–Planck problem $p \in C([0,T]; H^1(\Omega))$ and due to $q \in W(0,T)$, we obtain the regularity

$$q_0 \in H^1(\Omega) \quad \text{and} \quad f \in L^2(\Omega_T).$$

Hence, we can apply the regularity Lemma 2.3.1 to deduce that $q \in C([0,T]; H^1(\Omega)) \cap H^1(0,T; L^2(\Omega))$. The claim of (4.10) is therefore proven. The Lipschitz continuity of Θ follows analogously to the proof of Lemma 2.4.4 and using the fact that G is Lipschitz continuous. □

We can now prove the implicit representation of minimizer.

Proof of Theorem 4.1.5. The claim a) is a direct consequence of Lemma 1.4.6, which we apply in our case with $\Omega =]0,T[$ and $f = \Phi_i[\bar{u}] + \gamma \bar{u}_i$ for $i = 1, \dots, m$.
The claim b) is proven in a different manner, since no box–constraints are present. Thus, the first–order optimality conditions reads for $i = 1, \dots, m$

$$\langle \Phi_i[\bar{u}] + \gamma \bar{u}_i, v \rangle_2 + \langle \gamma \bar{u}_i', v' \rangle_2 = 0, \quad v \in H_0^1(0,T).$$

This is exactly the weak formulation of the following one–dimensional elliptic Dirichlet problem for w with right–hand–side $f = \frac{1}{\gamma}\Phi_i[\bar{u}]$ in strong form

$$w'' = w + f \qquad \text{on } [0,T], \qquad w(0) = 0 = w(T). \tag{4.13}$$

Obviously, $\Phi[\bar{u}] \in L^2(0,T)^m$, therefore by standard elliptic theory, the problem (4.13) possesses a unique strong solution $H^2(0,T) \cap H_0^1(0,T)$. Consequently, f.a.e. $t \in [0,T]$ it holds that

$$\bar{u}''(t) = \bar{u}(t) + \frac{1}{\gamma}\Phi[\bar{u}](t).$$

Since the r.h.s. is continuous, it must hold that u'' is continuous as well and the claim $u \in C^2([0,T])^m$ follows. □

For the proof of Theorem 4.1.6, we refer the reader to [5] and Theorem 1.3.6. This concludes the chapter, and we continue with a numerical analysis of this optimal control problem by a Galerkin discretization.

5 Numerical analysis for a class of optimal control problems with time–dependent controls

Calculus succeeds by breaking complicated problems down into simpler parts. That strategy, of course, is not unique to calculus. All good problem–solvers know that hard problems become easier when they're split into chunks. The truly radical and distinctive move of calculus is that it takes this divide-and-conquer strategy to its utmost extreme – all the way out to infinity.

STEVEN H. STROGATZ in *Infinite Powers: How Calculus Reveals the Secrets of the Universe*, 2019

It is the aim of Chapter 5–7 to establish accuracy estimates for the Fokker–Planck problem with a tracking type cost functional and only time–dependent control. This is a very challenging and complex task, which is why we have split it into three steps, and each step is performed in one chapter. First, we show under which condition a PDE optimal control problem can be approximated by an ODE–constrained optimal control problem, governed by a semidiscrete Galerkin approach. Secondly, we analyze this so–called semidiscrete optimization problem and establish accuracy estimates in Chapter 6. In the third step, performed in Chapter 7, we apply these results on our Fokker–Planck problem, which concludes the numerical analysis.

In this chapter, we perform the first step and present a framework for reducing the complexity of optimization problems of the following form

$$\min_{u\in U_{\text{ad}}} J(f,u) \quad \text{s.t.} \qquad \partial_t f + L[u](f) = 0, \quad f(0) = f_0, \tag{5.1}$$

where $f = f(t,x)$ represents the real–valued state function of space and time, and $\partial_t f$ denotes its partial time derivative. Furthermore, $f \mapsto L[u](f)$ denotes a differential operator with respect to x, acting only on its first argument f and including a control mechanism with the time–dependent control $u = u(t)$.

The optimal control problem (5.1) is defined on the spacetime cylinder $\Omega_T =]0, T[\times \Omega$, where $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ is convex, bounded, and is polygonal or has C^2–boundary, and $T > 0$ denotes the final time horizon as usual. We assume that appropriate boundary conditions are given on $\partial\Omega$, which are included in the definition of L. Additionally, f_0 represents the initial conditions. The set of admissible controls and the objective functional are denoted with U_{ad} and J, respectively. The control–to–state map $u \mapsto G(u) = f$ is assumed to be well–defined, and further properties for G and $\hat{J} := J(G(\cdot), \cdot)$ are given in the next section.

We remark that optimal control problems with only time–dependent control functions occur in many applications and research papers [6, 7, 16, 44, 51]. In many cases, the function u plays the role of a modulating function of a given space–depending potential.

Our main assumptions on (5.1) are second–order optimality results of $u \mapsto J(G(u), u)$ and convergence results of a semidiscrete Galerkin approximation of the PDE

$$\partial_t f + L[u](f) = 0, \quad f(0) = f_0 \tag{5.2}$$

for fixed u. We show then that establishing error estimates of (5.1) can be reduced to finding error estimates of an ODE–constrained optimization problem, where in many cases error estimates are available in the literature: In [30,62], Galerkin discretizations are presented for semilinear parabolic and hyperbolic PDEs. Second–order analysis for several optimal control problems can be found in [5, 21, 22, 46, 48, 64], and accuracy estimates for optimal control problems constrained by a system of (non)linear ODEs and their numerical approximation are analyzed in [3, 19, 23, 27, 29, 47, 60]. However, there are fewer results for accuracy estimates for PDE–constrained minimization problems, and with the theory presented in this chapter we would like to contribute to the research work on this field. For this purpose, we present a general approach that is based on the above–mentioned results and techniques in order to analyze PDE–constrained optimal control problems of the form given in (5.1).

Our approach is then applied in Chapter 7 to the Fokker–Planck optimal control problem with bilinear control mechanism of the form

$$\partial_t f - a\Delta f + \operatorname{div}\big(B[u]f\big) = 0,$$

where $B[u](t, x) = M(x)u(t) + c(x)$. To the best of the authors' knowledge, for this problem, no error analysis has been presented before. However, first– and second–order optimality conditions for this and similar bilinear problems have been established in [5,45,46], and accuracy estimates of the corresponding bilinear ODE–constrained problem can be found in [47] and are presented in Chapter 6.

This chapter is organized as follows. In the next section, we discuss the spatial finite element setting in detail. Furthermore, for a better understanding of our approach, the time discretization scheme that will be used in Chapter 7 for the semidiscrete problem is also introduced. Next, in Section 5.2, we formulate our approach for deriving accuracy estimates in a general framework. Section 5.3 is devoted to the numerical analysis of the Galerkin approximation of our Fokker–Planck problem.

5.1 The spatial discretization and the semidiscrete optimal control problem

We start this section by introducing our finite element setting for the spatial discretization. Let $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ be bounded and convex, and for simplicity in the notation of the discretization, we assume that Ω is a polygonal domain. Let $h > 0$ denote the discretization parameter and we introduce a quasi–uniform triangulation $\{\mathcal{T}_h\}_{h>0}$ of Ω with

$$\bar{\Omega} = \bigcup_{S\in\mathcal{T}_h} \bar{S} \quad \text{and} \quad h = \max\{\mathrm{diam}(S) : S \in \mathcal{T}_h\}.$$

Next, we introduce the $N = N(h)$–dimensional space of linear finite elements for $H^1(\Omega)$–functions given by

$$\mathcal{P}_\Omega^N := \left\{\psi \in C(\bar{\Omega}) : \psi \text{ is linear on each } S \in \mathcal{T}_h\right\}.$$

The canonical basis is denoted by $\{\psi_i \,:\, i = 1, \dots, N\}$ of $\mathcal{P}_\Omega^N$, where ψ_j is the typical pyramid function, see [15, Chapter 3]. Next, notice that for H^1– and H^2–functions in Ω and for the triangulation $\{\mathcal{T}_h\}_{h>0}$, the following accuracy result holds

$$\begin{aligned}
&\inf_{\psi\in\mathcal{P}_\Omega^N} \left\{\|f-\psi\|_{L^2(\Omega)}\right\} \le C_\Omega h \|f\|_{H^1(\Omega)}, \quad f \in H^1(\Omega),\\
&\inf_{\psi\in\mathcal{P}_\Omega^N} \left\{\|f-\psi\|_{L^2(\Omega)}\right\} + h \inf_{\psi\in\mathcal{P}_\Omega^N} \left\{\|f-\psi\|_{H^1(\Omega)}\right\} \le C_\Omega h^2 \|f\|_{H^2(\Omega)}, \quad f \in H^2(\Omega).
\end{aligned} \tag{5.3}$$

We recall the L^2–orthogonal projections $\mathrm{Proj}_{L^2(\Omega)}^N : L^2(\Omega) \to \mathcal{P}_\Omega^N$ defined by

$$\langle \mathrm{Proj}_{L^2(\Omega)}^N(f), \psi\rangle_{L^2(\Omega)} = \langle f, \psi\rangle_{L^2(\Omega)}, \quad \psi \in \mathcal{P}_\Omega^N. \tag{5.4}$$

Moreover, we introduce the Riesz–projection (or ∇–orthogonal projection) $\mathrm{Proj}_\nabla^N : H^1(\Omega) \to \mathcal{P}_\Omega^N$ defined by

$$\langle \nabla\mathrm{Proj}_\nabla^N(f), \nabla\psi\rangle_{L^2(\Omega)} = \langle \nabla f, \nabla\psi\rangle_{L^2(\Omega)}, \quad \psi \in \mathcal{P}_\Omega^N. \tag{5.5}$$

We recall important accuracy rates in the following Lemma; in view of (5.3), we see that these rates are optimal.

Lemma 5.1.1. *The following estimates hold for the L^2– and Riesz–projection. There exists $C_\Omega > 0$ such that for all mesh sizes $0 < h < 1$ and all $g \in H^1(\Omega)$, $f \in H^2(\Omega)$ we have*

$$\|g - \mathrm{Proj}_{L^2(\Omega)}^N(g)\|_{H^1(\Omega)')} + h\|g - \mathrm{Proj}_{L^2(\Omega)}^N(g)\|_{L^2(\Omega)} \le C_\Omega h^2 \|g\|_{H^1(\Omega)}, \tag{5.6}$$

$$\|f - \mathrm{Proj}_{L^2(\Omega)}^N(f)\|_{L^2(\Omega)} + h\|\nabla(f - \mathrm{Proj}_{L^2(\Omega)}^N(f))\|_{L^2(\Omega)} \le C_\Omega h^2 \|f\|_{H^2(\Omega)}, \tag{5.7}$$

$$\|f - \mathrm{Proj}_\nabla^N(f)\|_{L^2(\Omega)} + h\|\nabla(f - \mathrm{Proj}_\nabla^N(f))\|_{L^2(\Omega)} \le C_\Omega h^2 \|f\|_{H^2(\Omega)}, \tag{5.8}$$

Additionally, the following convergences hold for $g \in L^2(\Omega)$, $f \in H^1(\Omega)$

$$\|\mathrm{Proj}_{L^2(\Omega)}^N(g) - g\|_{L^2(\Omega)} + \|\mathrm{Proj}_\nabla^N(f) - f\|_{H^1(\Omega)} \to 0, \quad h \to 0. \tag{5.9}$$

Proof. A proof is given in [24], [62, Chapter 1] and in [33, Prop. 1.134] (where $l = 1$, $k = 1$), and we refer the reader to [15, Thm. 4.4.4] for analogous approximation properties in a more general setting. □

Let us sketch our strategy to discretize the optimal control problem (5.1) with the just presented finite element method. First, let $\mathcal{L}[u](f,\cdot)$ denote a weak formulation of $L[u](f)$ such that the classical solutions f satisfies for all test functions $\psi : \Omega \to \mathbb{R}$ the following

$$\begin{aligned}
&\langle \partial_t f, \psi\rangle_{L^2(\Omega)} + \mathcal{L}[u](f,\psi) = 0,\\
&\langle f_0 - f(0,\cdot), \psi\rangle_{L^2(\Omega)} = 0.
\end{aligned} \tag{5.10}$$

In a finite element approach, the ansatz is to seek semidiscrete solutions $f^h \in H^1(0,T;\mathcal{P}_\Omega^N)$ of the form

$$f^h(t,x) = \sum_{j=1}^N y_j(t)\psi_j(x).$$

Inserting this ansatz into the weak formulation (5.10) of f and replacing the space of test functions with the finite element space $\mathcal{P}_\Omega^N$, we obtain the following system of equations for $j = 1, \dots, N$

$$\langle \partial_t f^h, \psi_j \rangle_{L^2(\Omega)} + \mathcal{L}[u](f^h, \psi_j) = 0, \tag{5.11}$$

$$\langle f_0 - f^h(0,\cdot), \psi_j \rangle_{L^2(\Omega)} = 0. \tag{5.12}$$

Assuming the control u to be only time–dependent, we can carry out all integrations over x. Therefore, (5.11) becomes a system of ODEs for y with initial value given by (5.12). This justifies the terminology of finding f^h as a semidiscrete problem, since it is a discrete problem in x but still a Cauchy problem in t.

We proceed analogously for the semidiscretization of the objective J, and for a better illustration of our approach, we assume $J(f, u)$ to be of the form

$$J(f,u) = \int_{\Omega_T} \ell\big(t, x, f(t,x), u(t)\big)\, dx\, dt + \int_\Omega \kappa\big(x, f(T,x)\big)\, dx. \tag{5.13}$$

We replace all space dependent functions with its projections $\mathrm{Proj}^N_{L^2(\Omega)}(\cdot)$ to the finite element space, carry out the integrations over x and obtain a semidiscrete objective, that involves only time–integration of the form

$$J_h(y,u) = \int_0^T \ell_h\big(t, y(t), u(t)\big)\, dt + \kappa_h\big(y(T)\big).$$

In conclusion, we have – at least formally – obtained the ODE–constrained optimal control problem

$$\min_{u \in U_{\mathrm{ad}}} J_h(y,u) \quad y \text{ given by } \quad (5.11)\text{–}(5.12). \tag{5.14}$$

Under certain assumptions on the PDE (5.2) and the objective (5.13), one can expect that $f^h \to f$ in a suitable sense and $J_h(y,u) \to J(f,u)$ as the mesh size h tends to zero. Thus, we call (5.14) the semidiscrete optimal control problem, and we remark that both optimization problems are defined on the same set of admissible controls U_{ad}.

The idea of this splitting procedure is that, on one hand, we need to verify that this is a good approximation in the sense that the minimizers $\bar{u}_h$ of (5.14) converge to minimizers $\bar{u}$ of (5.1). In other words, we address the following question: If a sequence of cost functionals $\hat{J}_h$ converges to $\hat{J}$ with a certain rate, under which further conditions can we make a statement on the convergence rates of the local minimizers $\bar{u}_h$ to $\bar{u}$? This question will be answered in the next section.

On the other hand, we hope to have simplified the problem in the sense that (5.14) is a simple problem compared to (5.1) from a theoretical and numerical point of view. The simplified problem is analyzed in Chapter 6. Let us introduce its time discretization with uniform mesh size $k := T/K$, where $K \in \mathbb{N}$ is the number of grid points and $t_i := ik$ for $i = 0, \dots, K$. Let $(U_{\mathrm{ad}})_k$ be the corresponding finite–dimensional space to U_{ad} and let $\hat{J}_{h,k}$ be the finite–dimensional objective corresponding to $\hat{J}_h$. The corresponding finite–dimensional optimization problem reads

$$\min_{u \in (U_{\mathrm{ad}})_k} \hat{J}_{h,k}(u)$$

with solution $\bar{u}_{h,k} \in (U_{\mathrm{ad}})_k$. For a suitable projection $\mathcal{P}_k : (U_{\mathrm{ad}})_k \to U_{\mathrm{ad}}$, we have split the problem of establishing error estimates into

$$\|\bar{u} - \mathcal{P}_k(\bar{u}_{h,k})\|_2 \le \|\bar{u} - \bar{u}_h\|_2 + \|\bar{u}_h - \mathcal{P}_k(\bar{u}_{h,k})\|_2. \tag{5.15}$$

Let us put it in concrete terms for our Fokker–Planck problem. In that case, the semidiscrete problem for $\hat{J}_h$ is discretized in time with a finite element method, i.e., $(U_{\mathrm{ad}})_k$ is the space of piecewise constant

or piecewise quadratic polynomials and no projection is needed. Due to (5.15), we are able to prove first– or second–order accuracy, that is, minimizers $\bar{u}_{h,k}$ of $\hat{J}_{h,k}$ converge to $\bar{u}$ in $L^2(0,T)$ with rate $k^r + h^r$, where $r = 1$ or $r = 2$ depends on certain regularity assumptions.
The first term of (5.15) is treated in Section 5.2 and the second term is estimated in Chapter 6.

5.2 Accuracy estimates of optimization problems in an abstract framework

In this section, we formulate the conditions for the abstract minimization problem (5.1), that are based on second–order results from Section 1.3 and the splitting idea (5.15). We completely adopt the notation from Theorem 1.3.6. The set of admissible controls $\emptyset \neq U_{\text{ad}} \subset L^2(0,T)^m$, $m \in \mathbb{N}$, is convex and closed, and U is a Hilbert–space with scalar product $\langle \cdot, \cdot \rangle_U$ and norm $\| \cdot \|_U = \langle \cdot, \cdot \rangle_U^{1/2}$ that covers U_{ad}. Let U_∞ be a Banach space with norm $\| \cdot \|_\infty$ and continuous embedding $U_\infty \subset U$. Possible choices of U and U_∞ are, for example, $L^2(0,T), H_0^1(0,T)$ or $H^k(0,T)$ and $L^\infty(0,T)$, respectively; also the choice $U = U_\infty$ is possible.
Furthermore, we assume the existence of a control–to–state map $u \mapsto G(u)$, and introduce the reduced cost functional $\hat{J} : U \to \mathbb{R}$, $\hat{J}(u) := J(G(u), u)$. In this section, the minimization problem under consideration reads

$$\min_{u \in U_{\text{ad}}} \hat{J}(u). \tag{5.16}$$

We recall that $u^* \in U_{\text{ad}}$ is a minimizer of $\hat{J}$ (or a solution of (5.16)) if $\hat{J}(u^*) \leq \hat{J}(u)$ for all $u \in U_{\text{ad}}$. Furthermore, some $\bar{u} \in U_{\text{ad}}$ is a local minimum of $\hat{J}$ (or a local solution of (5.16)) if there exists $r > 0$ such that $\hat{J}(\bar{u}) \leq \hat{J}(u)$ for all $u \in U_{\text{ad}} \cap B_r(\bar{u}; U_\infty)$.
Let $h > 0$ denote the spatial mesh size for the space approximation. Let us fix $\bar{u} \in U_{\text{ad}}$ and recall the following conditions on $\hat{J}$ from Theorem 1.3.6:

(C1) $\hat{J}$, $\hat{J}_h$ is of class C^2 in U_∞ and for every $u \in U_{\text{ad}}$, there exists continuous extensions

$$\hat{J}'(u), \hat{J}_h'(u) \in \text{Lin}\,(U)\,, \quad \hat{J}''(u), \hat{J}_h''(u) \in \text{Bilin}\,(U \times U)\,. \tag{C1}$$

(C2) There exists $\Lambda > 0$ such that for all sequences $(u_n)_{n \in \mathbb{N}} \subset U_{\text{ad}}$ and $(v_n)_{n \in \mathbb{N}} \subset U$ with $u_n \to \bar{u}$ strongly in U and $v_n \rightharpoonup v$ weakly in U it holds that

$$\hat{J}'(\bar{u})v = \lim_{n \to \infty} \hat{J}'(u_n)v_n, \tag{C2.1}$$

$$\hat{J}''(\bar{u})(v,v) \leq \liminf_{n \to \infty} \hat{J}''(u_n)(v_n, v_n), \tag{C2.2}$$

$$\text{and if } v = 0, \text{ then } \quad \Lambda \liminf_{n \to \infty} \|v_n\|_U^2 \leq \liminf_{n \to \infty} \hat{J}''(u_n)(v_n, v_n). \tag{C2.3}$$

Furthermore, we state the following conditions for the semidiscrete problem.

(C3) The semidiscrete functional $\hat{J}_h : U_{\text{ad}} \to \mathbb{R}$ is well–defined and

$$\hat{J}_h(u) \to \hat{J}(u) \tag{C3}$$

as $h \to 0$, uniformly for $u \in U_{\text{ad}}$.

(C4) The sequence $(\bar{u}_h)$ of local minima of $\hat{J}_h$ exists and

$$\bar{u}_h \to \bar{u} \quad \text{ in } U \text{ as } h \to 0. \tag{C4}$$

(C5) For all $\delta > 0$ there exists $\varepsilon, h_0 > 0$ such that for all $v \in U$, $0 < h < h_0$ and $u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; U)$ it holds that

$$|\hat{J}_h''(u)(v,v) - \hat{J}''(u)(v,v)| \leq \delta \|v\|_U^2. \tag{C5}$$

(C6) Let $\tau > 0$. There exists some $h_0 > 0$ and rate $r > 0$ such that for $(\bar{u}_h)$ and $\bar{u}$ from (C4) it holds that

$$\left|\hat{J}'(\bar{u}_h)v - \hat{J}_h'(\bar{u}_h)v\right| \leq Ch^r \|v\|_U, \quad v \in U_{\mathrm{ad}}, \tag{C6}$$

for all $0 < h < h_0$, where $C > 0$ is independent of $\bar{u}_h$ and h. Additionally, $\bar{u}_h - \bar{u} \in E_{\bar{u}}^\tau$, where

$$E_{\bar{u}}^\tau = \{v \in \overline{S_{\bar{u}}}^U \ : \ |\hat{J}'(\bar{u})v| \leq \tau \|v\|_2\}.$$

In addition to these conditions, we require the following first– and second–order assumptions on $\bar{u}$

$$\hat{J}'(\bar{u})(u - \bar{u}) \geqslant 0, \quad u \in U_{\mathrm{ad}}, \tag{A1}$$

$$\hat{J}''(\bar{u})(v,v) > 0, \qquad v \in C_{\bar{u}} \backslash \{0\}. \tag{A2}$$

where $C_{\bar{u}}$ denotes the critical cone defined in Section 1.3.

We present the main result of our abstract approach in the following theorem. It specifies how the auxiliary term $\|\bar{u} - \bar{u}_h\|_U$ can be estimated. Therefore, it allows for the reduction of the complexity of the problem $\min_{u \in U_{\mathrm{ad}}} \hat{J}(u)$ to that of $\min_{u \in U_{\mathrm{ad}}} \hat{J}_h(u)$.

Theorem 5.2.1. *Let $\bar{u} \in U_{\mathrm{ad}}$ satisfy* (A1)–(A2) *from above, and let the conditions* (C1)–(C4) *and* (C6) *hold for the reduced objective functionals $\hat{J}, \hat{J}_h : U \to \mathbb{R}$ at $\bar{u}$.*
Then, for sufficiently small h, it holds that

$$\|\bar{u} - \bar{u}_h\|_U \leq Ch^r,$$

for some $C > 0$ independent of $\bar{u}_h$ and h.

Proof. We see that Theorem 1.3.6 implies local coercivity of $\hat{J}$ around $\bar{u}$ in the sense that there exists $\nu, \tau, \varepsilon > 0$ such that for all $u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; U_2)$ and all $v \in E_{\bar{u}}^\tau$, it holds that

$$\hat{J}''(u)(v,v) \geq \frac{\nu}{2} \|v\|_U^2.$$

Since $\bar{u}_h - \bar{u} \in E_{\bar{u}}^\tau$ by (C6), we have for all $h > 0$ and $w \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; U_2)$ that

$$\frac{\nu}{2} \|\bar{u} - \bar{u}_h\|_U^2 \leq J''(w)(\bar{u} - \bar{u}_h, \bar{u} - \bar{u}_h). \tag{5.17}$$

Due to the mean value theorem, for all $h > 0$ there exists $\lambda \in [0,1]$ such that

$$\hat{J}''(w_h)(\bar{u} - \bar{u}_h, \bar{u} - \bar{u}_h) = \left(\hat{J}'(\bar{u}_h) - \hat{J}'(\bar{u})\right)(\bar{u}_h - \bar{u}),$$

where $w_h := \lambda \bar{u} + (1 - \lambda)\bar{u}_h$. Due to the strong convergence $\bar{u}_h \to \bar{u}$ in U, we find $h_0 > 0$ such that for all $0 < h < h_0$ we have $w_h \in B_\varepsilon(\bar{u}; U)$. Since $-\hat{J}'(\bar{u})(\bar{u}_h - \bar{u}) \leq 0 \leq -\hat{J}_h'(\bar{u}_h)(\bar{u}_h - \bar{u})$ due to the FONC, it holds that

$$\begin{aligned}
\frac{\nu}{2} \|\bar{u} - \bar{u}_h\|_U^2 &\leq \hat{J}''(w_h)(\bar{u} - \bar{u}_h, \bar{u} - \bar{u}_h) \\
&= \left(\hat{J}'(\bar{u}_h) - \hat{J}'(\bar{u})\right)(\bar{u}_h - \bar{u}) \\
&\leq \left(\hat{J}'(\bar{u}_h) - \hat{J}_h'(\bar{u}_h)\right)(\bar{u}_h - \bar{u}) \\
&\leq Ch^r \|\bar{u}_h - \bar{u}\|_U,
\end{aligned}$$

where the last estimate is from (C6). Dividing by $\|\bar{u}_h - \bar{u}\|_U$ yields the claim b). □

Concerning (C4), in many applications a first–order necessary condition yields an implicit formula for $\bar{u}_h$ and $\bar{u}$, which can be exploited to derive the convergence of the norms $\|\bar{u}_h\|_U \to \|\bar{u}\|_U$. Another, more direct approach is shown in the following lemma, where a sequence of minima is constructed on a small closed ball $(\bar{u}_h)$ that converges strongly to $\bar{u}$.

Lemma 5.2.2. *Let $\hat{J}, \hat{J}_h$ be w.l.s.c on U_{ad} and fulfill* (C1)–(C3)*. Let* (A1)–(A2) *hold for $\bar{u} \in U_{\mathrm{ad}}$ and let $\varepsilon > 0$ be given by Theorem 5.2.1. Furthermore, let the objective $\hat{J}$ and the semidiscrete objective $\hat{J}_h$ be of the form*

$$\hat{J}(u) = F(u) + \frac{\gamma}{2}\|u\|_U^2, \qquad \hat{J}_h(u) = F_h(u) + \frac{\gamma}{2}\|u\|_U^2$$

for $\gamma > 0$ and functions $F, F_h : U \to \mathbb{R}$, and if $w_h \rightharpoonup w$ in U then $F_h(w_h) \to F(W)$. Then, the following holds:

a) There exists a sequence $(\bar{u}_h)_{h>0} \subset U_{\mathrm{ad}}$ of solutions to

$$\min\left\{\hat{J}_h(u) \mid u \in U_{\mathrm{ad}} \cap \overline{B_{\varepsilon/2}}^U(\bar{u};U)\right\}, \quad \overline{B_{\varepsilon/2}}^U(\bar{u};U) := \{u \in U \mid \|u - \bar{u}\|_U \le \varepsilon/2\}, \tag{5.18}$$

with $\bar{u}_h \rightharpoonup \bar{u}$ in U as $h \to 0$.

b) If $\|\bar{u}_h\|_U \to \|\bar{u}\|_U$, then $(\bar{u}_h)_{h>0}$ is a sequence of local minima of $\hat{J}_h$, and condition (C4) *is satisfied.*

c) The convergence of the norms $\|\bar{u}_h\|_U \to \|\bar{u}\|_U$ hold.

Proof. We start by showing the existence of solutions w_h to (5.18). Due to (C3), $\hat{J}_h$ is bounded from below, and we can pick a minimizing sequence $(w_h^n)_{n\in\mathbb{N}} \subset U_{\mathrm{ad}} \cap \overline{B_{\varepsilon/2}}^U(\bar{u};U)$ with

$$\lim_{n\to\infty} \hat{J}_h(w_h^n) = \inf\left\{\hat{J}_h(u) \mid u \in U_{\mathrm{ad}} \cap \overline{B_{\varepsilon/2}}^U(\bar{u};U)\right\}.$$

By construction, for every $h > 0$, the sequence $(w_h^n)_{n\in\mathbb{N}}$ is uniformly bounded in U by $\|\bar{u}\|_U + \varepsilon$, and hence, there exists a weak limit w_h in U (after selecting a subsequence). Due to the convexity and closedness of $U_{\mathrm{ad}} \cap \overline{B_{\varepsilon/2}}^U(\bar{u};U)$, we deduce by an application of Mazur's lemma, see Lemma 1.4.2, that $w_h \in U_{\mathrm{ad}} \cap \overline{B_{\varepsilon/2}}^U(\bar{u};U)$. Thus, the weak lower semicontinuity of $\hat{J}_h$ implies that w_h is a local minimizer. Next, we show that the sequence of minimizers $(w_h)_{h>0}$ converges weakly to $\bar{u}$ in U as $h \to 0$. Since $\|w_h\|_U \le \|\bar{u}\|_U + \varepsilon$, there exists a weak limit $w \in U_{\mathrm{ad}} \cap \overline{B_{\varepsilon/2}}^U(\bar{u};U)$. Indeed, we can prove $w = \bar{u}$ as follows. First, notice that the assumption on F_h, F implies $F_h(w_h) \to F(\bar{u})$. Hence, we obtain that

$$\hat{J}(w) \le \liminf_{h\to 0} \hat{J}_h(w_h) \quad \text{and} \quad \hat{J}_h(w_h) \le \hat{J}_h(\bar{u})$$

due to (5.18). Consequently, it holds that

$$\hat{J}(w) \le \liminf_{h\to 0} \hat{J}_h(w_h) \le \lim_{h\to 0} \hat{J}_h(\bar{u}) = \hat{J}(\bar{u}) \le \hat{J}(w). \tag{5.19}$$

Now recall the quadratic growth condition at $\bar{u}$ which implies that $\bar{u}$ is a strict local minimum, that is,

$$\hat{J}(\bar{u}) < \hat{J}(u), \quad u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u};U).$$

Since $\hat{J}(\bar{u}) = \hat{J}(w)$ and $w \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u};U)$, it follows that $w = \bar{u}$.

In conclusion, we have shown that a subsequence of (w_h) – let us redefine it as $(\bar{u}_h)$ – converges weakly to $\bar{u}$ in U as $h \to 0$, thereby proving part (a). Furthermore, if $\|\bar{u}_h\|_U \to \|\bar{u}\|_U$, this would imply strong convergence in U since U is a Hilbert space. Consequently, for sufficiently small h, every $\bar{u}_h$ has to be in the interior of the closed ball. Therefore, it is a local minimum of $\hat{J}_h$. This proves part b).

In order to verify assertion c), notice that in (5.19), we have proven that $\hat{J}(\bar{u}) = \liminf_{h\to 0} \hat{J}_h(\bar{u}_h)$, and hence,

$$\liminf_{h\to 0} \frac{\gamma}{2}\|\bar{u}_h\|_U^2 = \liminf_{h\to 0} \left(\hat{J}_h(\bar{u}_h) - F_h(\bar{u}_h)\right) = \hat{J}(\bar{u}) - F(\bar{u}) = \frac{\gamma}{2}\|\bar{u}\|_U^2.$$

Due to the weak convergence, it holds that $\bar{u}_h \to \bar{u}$ strongly in U for a subsequence. Since the limit $\bar{u}$ is unique for every subsequence, the entire sequence converges without the need for selecting a subsequence, see Lemma 1.4.3. This concludes the proof. □

Let us emphasize that, in general, for non–convex minimization problems, one cannot simply construct an auxiliary problem $\hat{J}_h$ and assume that (C3) implies that its minimizers $\bar{u}_h$ converge to the minimizer $\bar{u}$ of $\hat{J}$. This is due to the fact that minimizers need not be unique, and local minima need not be strict. However, Lemma 5.2.2 ensures that for given $\bar{u}$, one picks the "correct" sequence of local minimizers $\bar{u}_h$ of $\hat{J}_h$, and this is where the second–order assumption (A2) for $\bar{u}$ enters in our approach.

Furthermore, second–order conditions on local minimizers are important for the numerical analysis of the auxiliary problem and for establishing accuracy estimates. Since the minimizers $\bar{u}_h$ are given by Lemma 5.2.2, one has to verify that the second–order condition of $\bar{u}$ will be passed on to $\bar{u}_h$. This can be done as follows. In the case of no box constraints on the controls, that is $U_{\text{ad}} = U$, the critical cone from (A2) becomes $C_{\bar{u}} = U$. Hence, if additionally condition (C5) holds, we can ensure that also $\bar{u}_h$ fulfills a second–order condition for $\hat{J}_h$.

In the case of $U_{\text{ad}} \subsetneq U$, the situation becomes more delicate, since we have to compare the critical cones of $\hat{J}$ and $\hat{J}_h$. In Section 7.2, where we investigate the non–convex Fokker–Planck problem, we solve this problem by defining two extended critical cones. This allows us with (C5) to deduce local coercivity of $\hat{J}_h$ around $\bar{u}_h$ from the given local coercivity of $\hat{J}$ around $\bar{u}$.

This concludes our discussion of the abstract minimization problem, and we focus on our Fokker–Planck optimal control problem of tracking type.

5.3 Convergence of the semidiscrete Galerkin approximation for the Fokker–Planck problem

In the following, we analyze the spatial Galerkin approximation to the FP problem. One difficulty is that the objective functional J contains the term $p(T,\cdot)$, and consequently, we need to analyze accuracy additionally in the $L^\infty(0,T;L^2(\Omega))$–norm. In order to obtain convergence of the semidiscrete Galerkin scheme, the $H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H^2(\Omega))$–regularity of solutions to the FP problem and its adjoint are sufficient. However, in order to derive quadratic accuracy of the semidiscrete Galerkin scheme, we have to assume a $L^2(0,T;H^3(\Omega)) \cap C([0,T];H^2(\Omega))$–regularity, at least for solutions of the form $\bar{p} = G(\bar{u})$, $\bar{\varrho} = \Theta(\bar{u})$, where $\bar{u} \in U_{\text{ad}}^T$ is a local minimum on $\hat{J}$ or $\hat{J}_h$. Let us comment on why it is reasonable to assume that such regular solutions $\bar{p}, \bar{\varrho}$ exist. The aim is to derive accuracy results for local minimizer $\bar{u}$ of $\hat{J}$. As we have seen in Theorem 4.1.5 a), the implicit representation of $\bar{u}$ allows obtaining higher regularity for local minimizer; in this case we have improved $\bar{u}$ from $L^\infty(0,T)$ to $H^{1/2}(0,T)$. Since $\bar{u}$ appears on the r.h.s. of the inhomogeneous heat equation for $\bar{p}$, see Lemma 2.3.2, possibly, $\bar{p}$ has now higher regularity than $H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H^2(\Omega))$, and we can repeat this argument. This bootstrap argument is known to work well for linear optimal control problems, however, since the FP problem is a bilinear problem, it is still ongoing work to rigorously verify this argument.

Next, recall the linear finite element setting on $H^1(\Omega)$ introduced in Section 5.1, where the set of pyramid functions $\{\psi_i : i = 1,\dots,N\}$ is the basis of $\mathcal{P}_\Omega^N$. The mass matrix is denoted by $\mathcal{M} \in \mathbb{R}^{N\times N}$ and has

ij–th entry

$$\mathcal{M}_{ij} := \langle \psi_i, \psi_j \rangle_{L^2(\Omega)}. \tag{5.20}$$

We notice that $\mathcal{M}$ is symmetric positive definite and has full rank, and therefore, we can define the two equivalent norms on $L^2(0,T)^N$ and $\mathbb{R}^N$, respectively,

$$\|y\|_{2,\mathcal{M}} := \left(\int_0^T y(t)^\top \mathcal{M} y(t)\, dt \right)^{1/2} \quad \text{for } y \in L^2(0,T)^N, \qquad |a|_{\mathcal{M}} := \left(a^\top \mathcal{M} a\right)^{1/2} \quad \text{for } a \in \mathbb{R}^N.$$

We are interested in solutions $P = P_N \in H^1(0,T; H^1(\Omega))$ of the form

$$P_N(t,x) = \sum_{i=1}^N y_i(t)\psi_i(x), \tag{5.21}$$

such that f.a.e. $t \in]0,T[$ and for $i = 1,\dots,N$ it holds

$$\langle \dot{P}_N(t), \psi_i \rangle_{L^2(\Omega)} + \mathcal{F}_t(P_N(t), \psi_i) = 0, \tag{5.22}$$

$$\langle P_N(0), \psi_i \rangle_{L^2(\Omega)} = \langle p_0, \psi_i \rangle_{L^2(\Omega)}. \tag{5.23}$$

This is equivalent to $y \in H^1(0,T)^N$ solving the linear system of differential equations given by

$$\mathcal{M}\, y'(t) = \left(\tilde{A} + u(t)\tilde{B}\right) y(t), \quad t \in [0,T] \tag{5.24}$$

with initial condition $y_i(0) = \langle p_0, \psi_i \rangle_2$ for $i = 1,\dots,N$. This is due to the fact that for the ij–th components, we have $\left(\tilde{A} + u(t)\tilde{B}\right)_{ij} = -\mathcal{F}_t[u](\psi_j, \psi_i)$ for $u \in L^2(0,T)^m$, and hence, the entries for $i,j = 1,\dots,N$ are given by

$$\tilde{A}_{ij} = -\int_\Omega \left(a\nabla\psi_j(x) - \psi_j(x)c(x)\right) \cdot \nabla\psi_i(x)\, dx,$$

$$(u(t)\tilde{B})_{ij} = -\int_\Omega \psi_j(x)\, \nabla\psi_i(x)^\top M(x)\, u(t)\, dx.$$

Since $\mathcal{M}$ is invertible, we can bring this problem in a standard ODE form by applying $\mathcal{M}^{-1}$ from the left on the equation. If $m = 1$, we define $A := \mathcal{M}^{-1}\tilde{A}$ and $B := \mathcal{M}^{-1}\tilde{B}$, and thus, (5.24) becomes $y' = (A + uB)y$.

By the Carathéodory theorem, there exists a unique solution $y \in H^1(0,T)^N$ to (5.24) and we refer to Chapter 6 for more details. For every $N \in \mathbb{N}$, this gives rise to the following unique semidiscretized control–to–state map

$$Y_N : L^2(0,T) \to H^1(0,T)^N, \quad u \mapsto y, \quad \text{or equivalently} \tag{5.25}$$

$$G_N : L^2(0,T) \to H^1(0,T; \mathcal{P}_\Omega^N), \quad u \mapsto P_N = \sum_{i=1}^N y_i\, \psi_i. \tag{5.26}$$

Now, we can state the main result of this section. For that purpose, we recall the following notations for the norms of Banach valued Lebesgue spaces:

$$\|\cdot\|_{L^pL^q} := \|\cdot\|_{L^p(0,T;L^q(\Omega))}, \quad \|\cdot\|_{L^pH^q} := \|\cdot\|_{L^p(0,T;H^q(\Omega))}, \quad \text{for } p,q \in [1,\infty].$$

Furthermore, let us recall the set of admissible controls under consideration

$$U_{\text{ad}}^T = \{u \in L^\infty(0,T)^m \mid u^{\min} \leq u_i(t) \leq u^{\max} \quad \text{f.a.e. } t \in [0,T],\, i = 1,\dots,m\}.$$

Theorem 5.3.1. *Let $u \in U_{\mathrm{ad}}^T$, $p = G(u) \in H^1(0,T;L^2(\Omega)) \cap C([0,T];H^1(\Omega))$ and $P = P_N \in H^1(0,T;\mathcal{P}_\Omega^N)$ be solutions to the FP problems* (2.10) *and* (5.22)*, respectively. Then, for any $\Phi \in H^1(0,T;\mathcal{P}_\Omega^N)$ of the form* (5.21)*, there exists a constant $C = C_{\mathrm{ad}} C_{\mathrm{F}*} > 0$ such that*

$$\|p - P\|_{L^\infty L^2} + \|p - P\|_{L^2 H^1} \leq C\big(\|\dot{p} - \dot{\Phi}\|_{L^2(\Omega_T)} + \|p - \Phi\|_{L^\infty L^2} + \|p - \Phi\|_{L^2 H^1}\big). \tag{5.27}$$

Proof. Let us define $w := p - P \in C([0,T];H^1(\Omega))$. Note that by the Galerkin–orthogonality, $w(t) \in H^1(\Omega)$ is a valid test function for the FP equation (2.10) but not for the semidiscretized one (5.22) since $p(t) \notin \mathcal{P}_\Omega^N$ in general. Thus, we obtain for any $\Phi \in H^1(0,T;\mathcal{P}_\Omega^N)$ of the form (5.21) the additional terms

$$\langle \dot{w}(t), w(t)\rangle_{L^2(\Omega)} + \mathcal{F}_t(w(t), w(t)) = \langle \dot{w}(t), p(t) - \Phi(t)\rangle_{L^2(\Omega)} + \mathcal{F}_t(w(t), p(t) - \Phi(t)), \quad \text{f.a.e. } t \in]0,T[\,.$$

Next, we recall Lemma 2.1.2 and deduce the existence of constants $\beta, \tilde{\gamma} > 0$ such that

$$-\mathcal{F}_t(\varphi,\varphi) \leq \tilde{\gamma}\|\varphi\|_{L^2(\Omega)}^2 - \beta\|\varphi\|_{H^1(\Omega)}^2$$

for all $\varphi \in H^1(\Omega)$. Thus, integrating with respect to t yields

$$\begin{aligned} \frac{1}{2}\|w(t)\|_{L^2(\Omega)}^2 \leq & \frac{1}{2}\|w(0)\|_{L^2(\Omega)}^2 + \int_0^t \Big(\tilde{\gamma}\|w(s)\|_{L^2}^2 - \beta\|w(s)\|_{H^1}^2\Big)\, ds & (5.28)\\ & + \int_0^t \langle \dot{w}(s), p(s) - \Phi(s)\rangle_{L^2(\Omega)}\, ds + \int_0^t \mathcal{F}_s(w(s), p(s) - \Phi(s))\, ds. & (5.29)\end{aligned}$$

We estimate the first term in line (5.29) by partial integration, L^2–Hölder–inequality and the ε–Young–inequality to obtain f.a.e. $t \in]0,T[$ and every small $\varepsilon > 0$ the following estimate

$$\begin{aligned} &\int_0^t \langle \dot{w}(s), p(s) - \Phi(s)\rangle_{L^2(\Omega)}\, ds \\ &\quad\leq \int_0^t \|w(s)\|_{L^2(\Omega)}\|\partial_t(p-\Phi)(s)\|_{L^2(\Omega)}\, ds \\ &\qquad + \|w(t)\|_{L^2(\Omega)}\|p(t) - \Phi(t)\|_{L^2(\Omega)} + \|w(0)\|_{L^2(\Omega)}\|p(0) - \Phi(0)\|_{L^2(\Omega)} \\ &\quad\leq \frac{1}{2}\int_0^t \|w(s)\|_{L^2(\Omega)}^2\, ds + \frac{1}{2}\|\partial_t(p-\Phi)\|_{L^2(\Omega_T)}^2 \\ &\qquad + \varepsilon\|w(t)\|_{L^2(\Omega)}^2 + C_\varepsilon\|p(t) - \Phi(t)\|_{L^2(\Omega)}^2 + \frac{1}{2}\|w(0)\|_{L^2(\Omega)}^2 + \frac{1}{2}\|p(0) - \Phi(0)\|_{L^2(\Omega)}^2. \end{aligned} \tag{5.30}$$

For the last integral in (5.29), the boundedness from Lemma 2.1.2 of the flux $\mathcal{F}$ similarly yields

$$\int_0^t \mathcal{F}_s(w(s), p(s) - \Phi(s))\, ds \leq \varepsilon\|w\|_{L^2(0,t;H^1(\Omega))}^2 + C_\varepsilon\|p - \Phi\|_{L^2H^1}^2.$$

Combining both estimates, we find that

$$\begin{aligned} \left(\frac{1}{2} - \varepsilon\right)\|w(t)\|_{L^2(\Omega)}^2 \leq & \|w(0)\|_{L^2(\Omega)}^2 + C_\varepsilon\|p - \Phi\|_{L^\infty L^2}^2 - (\beta - \varepsilon)\|w\|_{L^2(0,t;H^1(\Omega))}^2 \\ & + C_\varepsilon\|p - \Phi\|_{L^2H1}^2 + \frac{1}{2}\|\partial_t(p-\Phi)\|_{L^2(\Omega_T)}^2 + \int_0^t \left(\gamma + \frac{1}{2}\right)\|w(s)\|_{L^2(\Omega)}^2\, ds. \end{aligned} \tag{5.31}$$

In conclusion, Grönwall's lemma now implies that

$$\|w(t)\|_{L^2(\Omega)}^2 + \|w\|_{L^2H^1}^2 \leq C\Big(\|w(0)\|_{L^2(\Omega)}^2 + \|p - \Phi\|_{L^\infty L^2}^2 + \|p - \Phi\|_{L^2H^1}^2 + \|\partial_t(p-\Phi)\|_{L^2(\Omega_T)}^2\Big)$$

Lastly, we can estimate $w(0)$ due to (5.23) as follows

$$\|w(0)\|_{L^2(\Omega)} \leq \|p_0 - \Phi(0)\|_{L^2(\Omega)} \leq \|p - \Phi\|_{L^\infty L^2}.$$

This concludes the proof. □

We will see in Theorem 5.3.3 below that under further assumptions on Φ, the first term in (5.30) can be estimated in another way. As a consequence of Theorem 5.3.1, we get the following error estimate.

Corollary 5.3.2. *(Linear and quadratic error estimates for the semidiscrete Galerkin approximation) Let $r \in \{1,2\}$ and assume that $p = G(u) \in L^2(0,T;H^{1+r}(\Omega)) \cap H^1(0,T;H^r(\Omega))$. Then, there exists a constant $C = C_{\text{ad}}C_{\text{F}*}$ such that*

$$\|p - P\|_W \leq Ch^r\big(\|p\|_{L^2H^{1+r}} + \|p\|_{L^\infty H^r} + \|\partial_t p\|_{L^2H^r}\big). \tag{5.32}$$

Proof. Let us consider the case $r = 1$. Since $p \in L^2(0,T;H^2(\Omega)) \cap H^1(0,T;H^1(\Omega))$, we can define $\Phi(t) := \text{Proj}_\nabla^N(p(t)) \in \mathcal{P}_\Omega^N$ for $t \in [0,T]$, and due to the accuracy results from Lemma 5.1.1, it holds that

$$\begin{aligned}
&\|p(t) - \Phi(t)\|_{L^2(\Omega)} \leq Ch\|p(t)\|_{H^1} \leq Ch\|p\|_{L^\infty H^1},\\
&\|p - \Phi\|_{L^2H^1} = \left(\int_0^T \|p(t) - \Phi(t)\|_{H^1}^2\,dt\right)^{1/2} \leq Ch\|p\|_{L^2H^2},\\
&\|\partial_t(p - \Phi)\|_{L^2(\Omega_T)} \leq Ch\|\partial_t p\|_{L^2H^1} \leq Ch\|p\|_{H^1H^1}.
\end{aligned}$$

Notice that for the last estimate, we used that $\partial_t\Phi(t) = \text{Proj}_\nabla^N(\partial_t p(t))$ f.a.e. $t \in [0,T]$. An application of Theorem 5.3.1 yields the desired linear accuracy rate. This proves the claim for the linear accuracy estimate $r = 1$.
If $r = 2$, we need the regularity $p \in H^1(0,T;H^2(\Omega)) \cap L^2(0,T;H^3(\Omega))$, and the assertion follows analogously. □

As mentioned before, let us go through the proof of Theorem 5.3.1 in a different way and exploit the L^2–orthogonality of the L^2–projection $\text{Proj}_{L^2(\Omega)}^N$. Compared to Corollary 5.3.2, lower regularity of p is necessary for linear error estimates.

Theorem 5.3.3. *(Improved linear error estimates)*
Let $u \in U_{\text{ad}}^T$, $p = G(u) \in H^1(0,T;L^2(\Omega)) \cap L^2(0;H^2(\Omega))$ and $P = P_N \in H^1(0,T;\mathcal{P}_\Omega^N)$ be solutions to the FP problems (2.10) and (5.22), respectively, where $h > 0$ denotes the mesh size, and $N = N(h)$ is the dimension of the finite element space.

a) Then, there exists $C = C_{\text{ad}}C_{\text{F}*}$ such that

$$\|p - P\|_{L^\infty L^2} + \|p - P\|_{L^2H^1} \leq Ch(\|p\|_{L^2H^2} + \|p\|_{H^1L^2}).$$

b) Let $f \in L^2(\Omega_T)$. Let $z \in H^1(0,T;L^2(\Omega)) \cap L^2(0;H^2(\Omega))$ and $Z = Z_N \in H^1(0,T;\mathcal{P}_\Omega^N)$ be solutions to the inhomogeneous FP problems f.a.e. $t \in [0,T]$

$$\langle \dot{z}(t), \psi\rangle_{L^2(\Omega)} + \mathcal{F}_t(z(t),\psi) = \langle f,\psi\rangle_{L^2(\Omega)}, \quad \psi \in H^1(\Omega_T) \tag{5.33}$$

$$\langle \dot{Z}_N(t), \psi_i\rangle_{L^2(\Omega)} + \mathcal{F}_t(Z_N(t),\psi_i) = \langle f,\psi_i\rangle_{L^2(\Omega)}, \quad i = 1,\dots,N, \tag{5.34}$$

with initial condition $z(0,\cdot) = 0$ and $Z_N(0,\cdot) = 0$ a.e. on Ω. Then, the same linear accuracy estimate holds for $C = C_{\text{ad}}C_{\text{F}*}$

$$\|z - Z\|_{L^\infty L^2} + \|z - Z\|_{L^2H^1} \leq Ch(\|z\|_{L^2H^2} + \|z\|_{H^1L^2}).$$

c) Let $\varrho = \Theta(u)$ be the adjoint from Definition 4.1.4 and let $\varrho_N \in H^1(0,T;\mathcal{P}_\Omega^N)$ denote its Galerkin approximation, i.e. it satisfies

$$\begin{aligned}
-\langle \dot{\varrho}_N(t), \psi_i\rangle_{L^2(\Omega)} + \mathcal{F}_t[u](\psi_i, \varrho_N(t)) &= \beta\langle p(t) - p^d(t), \psi_i\rangle_{L^2(\Omega)}\\
\langle \varrho_N(T), \psi_i\rangle_{L^2(\Omega)} &= \alpha\langle p(T) - p^T, \psi_i\rangle_{L^2(\Omega)}
\end{aligned}$$

f.a.e. $t \in [0,T]$ *and* $i = 1, \dots, N$. *Then, a linear accuracy estimate holds with* $C = C_{\mathrm{ad}} C_{\mathrm{F}*} C_J$

$$\|\varrho - \varrho_N\|_{L^\infty L^2} + \|\varrho - \varrho_N\|_{L^2 H^1} \leq Ch(\|\varrho\|_{L^2H^2} + \|\varrho\|_{H^1L^2}).$$

d) *If the norms for* p, z, ϱ *on the r.h.s of the above estimates* $\|\cdot\|_{L^2H^2}$ *and* $\|\cdot\|_{H^1L^2}$ *are replaced by* $\|\cdot\|_{L^2H^3}$ *and* $\|\cdot\|_{H^1H^1}$, *then the linear rate* h *can be replaced by the quadratic rate* h^2.

e) *The Galerkin scheme converges weakly in* $H^1(0,T;L^2(\Omega))$ *and weakly* in* $C([0,T];H^1(\Omega))$, *uniformly on* U_{ad}^T, *i.e.,*

$$\begin{aligned} P_N \rightharpoonup^* p, \quad \varrho_N \rightharpoonup^* \varrho \qquad & \text{in } C([0,T];H^1(\Omega)) \quad \text{as } N \to \infty, \\ P_N \rightharpoonup p, \quad \varrho_N \rightharpoonup \varrho \qquad & \text{in } H^1(0,T;L^2(\Omega)) \quad \text{as } N \to \infty, \end{aligned}$$

Proof. Let us start with assertion a). We repeat the proof of Theorem 5.3.1 with $w := p - P$ until we arrive at (5.30). This time, we choose the L^2–projection $\Phi(t) := \mathrm{Proj}^N_{L^2(\Omega)}(p(t)) \in \mathcal{P}^N_\Omega$ f.a.e. $t \in [0,T]$ which implies the regularity

$$\Phi \in L^2(0,T;H^1(\Omega)) \cap H^1(0,T;L^2(\Omega)).$$

We exploit the L^2–orthogonality (5.4) and obtain f.a.e. $s \in [0,T]$ the following identity

$$\begin{aligned} \langle \dot{w}(s), p(s) - \Phi(s)\rangle_{L^2(\Omega)} &= \langle \dot{p}(s), p(s) - \Phi(s)\rangle_{L^2(\Omega)} \\ &= \langle \dot{p}(s) - \dot{\Phi}(s), p(s) - \Phi(s)\rangle_{L^2(\Omega)} \\ &= \frac{1}{2}\frac{d}{ds}\left(\|p(s) - \Phi(s)\|^2_{L^2(\Omega)}\right). \end{aligned}$$

Consequently, (5.30) becomes f.a.e. $t \in [0,T]$

$$\int_0^t \langle \dot{w}(s), p(s) - \Phi(s)\rangle_{L^2(\Omega)}\, ds = \frac{1}{2}\left(\|p(t) - \Phi(t)\|^2_{L^2(\Omega)} - \|p(0) - \Phi(0)\|^2_{L^2(\Omega)}\right).$$

We estimate the second term in (5.29) as before. Next, we insert our estimates in (5.28)–(5.29) and obtain that

$$\begin{aligned} \frac{1}{2}\|w(t)\|^2_{L^2(\Omega)} \leq & \frac{1}{2}\|w(0)\|^2_{L^2(\Omega)} + \int_0^t \left(\tilde{\gamma}\|w(s)\|^2_{L^2} - \beta\|w(s)\|^2_{H^1}\right) ds \\ & + \frac{1}{2}\|p(t) - \Phi(t)\|^2_{L^2(\Omega)} - \frac{1}{2}\|p(0) - \Phi(0)\|^2_{L^2(\Omega)} \\ & + \varepsilon\|w\|^2_{L^2(0,t;H^1(\Omega))} + C_\varepsilon\|p - \Phi\|^2_{L^2H^1}. \end{aligned}$$

Taking $\varepsilon = \beta$ and applying Grönwall's lemma yields the desired estimate for $\sup_{t\in[0,T]}\|w(t)\|_{L^2(\Omega)}$. Furthermore, taking $\varepsilon < \beta$, rearranging the estimate and estimating $\|w(0)\|_{L^2(\Omega)} \leq C\|p - \Phi\|_{L^\infty L^2}$ implies the desired estimate for w in the $L^2(0,T;H^1(\Omega))$–norm, and we arrive at

$$\|w\|^2_{L^\infty L^2} + \|w\|^2_{L^2H^1} \leq C\Big(\|p - \Phi\|^2_{L^\infty L^2} + \|p - \Phi\|^2_{L^2H^1}\Big). \tag{5.35}$$

Lastly, we apply the accuracy estimates from Lemma 5.1.1 to obtain the linear rates

$$\begin{aligned} &\|p - \Phi\|_{L^\infty L^2} = \|p - \Phi\|_{C([0,T];L^2(\Omega))} \leq C_\Omega h\|p\|_{C([0,T];H^1(\Omega))} \\ &\|p - \Phi\|_{L^2H^1} \leq C_\Omega h\|p\|_{L^2H^2}. \end{aligned} \tag{5.36}$$

This concludes the proof of a).

Assertion b) is proven analogously for $w := z - Z$, since we observe that in the first step the r.h.s. f cancels.

In order to show claim c), we define $w := \varrho - \varrho_N$ and $\Phi(t) := \mathrm{Proj}^N_{L^2(\Omega)}(\varrho(t)) \in \mathcal{P}^N_\Omega$. We notice that this problem is similar to problem b), where the bilinear flux operator $\mathcal{F}$ is replaced with its adjoint operator. We may repeat the proof of a) and obtain with (2.1.2) the estimate

$$\begin{aligned}\frac{1}{2}\|w(t)\|^2_{L^2(\Omega)} \leq \frac{1}{2}\|w(T)\|^2_{L^2(\Omega)} + \int_t^T \left(\tilde{\gamma}\|w(s)\|^2_{L^2(\Omega)} - \beta\|w(s)\|^2_{H^1(\Omega)}\right) ds \\ + \int_t^T \mathcal{F}_s(\varrho(s) - \Phi(s), w(s))\, ds.\end{aligned}$$

The last term is estimated again with Lemma 2.1.2 and the ε-Young inequality

$$\int_t^T \mathcal{F}_s(\varrho(s) - \Phi(s), w(s))\, ds. \leq \varepsilon \int_t^T \|w(s)\|^2_{H^1(\Omega)}\, ds + C_\varepsilon \|\varrho - \Phi\|^2_{L^2H^1}.$$

The other terms are treated as in part a). This completes the proof of c).
In order to prove part d), consider the estimate (5.35). This time, we apply estimate (5.7) from Lemma 5.1.1 to obtain the quadratic rates. For that purpose, let $f = p, z$ or ϱ satisfy the higher regularity, and replace (5.36) by

$$\begin{aligned}\|f - \Phi\|_{L^\infty L^2} &\leq C_\Omega h^2 \|f\|_{L^\infty H^2}, \\ \|f - \Phi\|_{L^2H^1} &\leq C_\Omega h^2 \|f\|_{L^2H^3}.\end{aligned}$$

This concludes the proof of d). In order to prove e) for the FP problem, we test the finite element formulation with $\dot{P} \in L^2(0,T;\mathcal{P}^N_\Omega)$ (instead of P) to obtain a.e. on $[0,T]$

$$\langle \dot{P}, \dot{P}\rangle_{L^2(\Omega)} + \mathcal{F}(P, \dot{P}) = 0. \tag{5.37}$$

Furthermore, it holds a.e. on $[0,T]$

$$\begin{aligned}\mathcal{F}(P,\dot{P}) &= a\langle \nabla P, \nabla \dot{P}\rangle_{L^2(\Omega)} + \langle P, B[u]\nabla \dot{P}\rangle_{L^2(\Omega)} \\ &= \frac{a}{2}\frac{d}{dt}\|\nabla P\|^2_{L^2(\Omega)} + \langle \mathrm{div}\,(P\,B[u]), \dot{P}\rangle_{L^2(\Omega)}.\end{aligned} \tag{5.38}$$

Integrating (5.37) w.r.t. the time variable and inserting (5.38) yields f.a.e. $t \in [0,T]$

$$0 = \int_0^t \left(\|\dot{P}(s)\|^2_{L^2(\Omega)} + \langle \mathrm{div}\,(P(s)B[u]), \dot{P}(s)\rangle_{L^2(\Omega)}\right) ds + \frac{a}{2}\left(\|\nabla P(t)\|^2_{L^2(\Omega)} - \|\nabla P(0)\|^2_{L^2(\Omega)}\right).$$

Rearranging this equation, applying Cauchy-Schwarz inequality and then an ε-Young inequality implies the estimate

$$\begin{aligned}&\frac{a}{2}\|\nabla P(t)\|^2_{L^2(\Omega)} + \int_0^t \|\dot{P}(s)\|^2_{L^2(\Omega)}\, ds \\ &\quad\leq \|\nabla P(0)\|^2_{L^2(\Omega)} + \int_0^t \|P(s)\|_{H^1(\Omega)} \|B[u](s)\|_{W^{1,\infty}(\Omega)} \|\dot{P}(s)\|_{L^2(\Omega)}\, ds \\ &\quad\leq \|\nabla P(0)\|^2_{L^2(\Omega)} + C_\varepsilon \int_0^t \|P(s)\|^2_{H^1(\Omega)}\, ds + \|B[u]\|^2_{L^\infty(0,T;W^{1,\infty}(\Omega))}\, \varepsilon \int_0^t \|\dot{P}(s)\|^2_{L^2(\Omega)}\, ds.\end{aligned}$$

Next, we exploit the uniform boundedness $\|B[u]\|^2_{L^\infty(0,T;W^{1,\infty}(\Omega))} \leq C_{\mathrm{ad}} C_{\mathrm{F}*}$, and therefore, we can choose $\varepsilon > 0$ sufficiently small such that

$$\frac{a}{2}\|\nabla P(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\int_0^t \|\dot{P}(s)\|^2_{L^2(\Omega)}\, ds \leq \|\nabla P(0)\|^2_{L^2(\Omega)} + C_\varepsilon \int_0^t \|P(s)\|^2_{H^1(\Omega)}\, ds.$$

Since this estimate holds f.a.e. $t \in [0,T]$, we can take the supremum over $[0,T]$. Furthermore, we estimate P in the $L^2(0,T;H^1(\Omega))$–norm by part a), and due to the equation (5.23) for $P(0)$, it holds that

$$\|\nabla P(0)\|^2_{L^2(\Omega)} = \left\|\nabla\left(\mathrm{Proj}^N_{L^2(\Omega)}(p(0))\right)\right\|^2_{L^2(\Omega)} \leq C\|p_0\|_{H^1(\Omega)}.$$

This implies

$$\sup_{t\in[0,T]} \frac{a}{2}\|\nabla P(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|\dot{P}\|_{L^2(0,T;L^2(\Omega))} \leq C(\|p_0\|_{H^1(\Omega)}\|P\|_{L^2(0,T;H^1(\Omega))} \leq C_{\mathrm{ad}}C_{\mathrm{F}*}.$$

Hence, the semidiscrete solutions P are bounded in $C([0,T];H^1(\Omega))$ and $H^1(0,T;L^2(\Omega))$, uniformly in N. Since the L^2–limit of P is p, the corresponding weak (or weak*) convergence holds without a selection of a subsequence and its limit has to be p which can be shown by a standard argument as in (2.46). The case for the adjoint is done analogously. □

Let us mention, that similarly to e) and a), we can derive the strong convergence in $C([0,T];H^1(\Omega))$ and $H^1(0,T;L^2(\Omega))$ by testing with $\dot{w}$ instead of w. Let $\Phi(t) := \mathrm{Proj}^N_{L^2(\Omega)}(\varrho(t)) \in \mathcal{P}^N_\Omega$ and observe that f.a.e. $t \in [0,T]$

$$\langle \dot{w}(t), \dot{w}(t)\rangle_{L^2(\Omega)} + \mathcal{F}_t(w(t), \dot{w}(t)) = \langle \dot{w}(t), \dot{p}(t) - \dot{\Phi}(t)\rangle_{L^2(\Omega)} + \mathcal{F}_t(w(t), \dot{p}(t) - \dot{\Phi}(t)), \tag{5.39}$$

Due to the L^2–orthogonality, we have this time f.a.e. $s \in [0,T]$

$$\begin{aligned}\langle \dot{w}(s), \dot{p}(s) - \dot{\Phi}(s)\rangle_{L^2(\Omega)} &= \langle \dot{p}(s) - \dot{\Phi}(s), \dot{p}(s) - \dot{\Phi}(s)\rangle_{L^2(\Omega)} \\ &= \|\dot{p}(s) - \dot{\Phi}(s)\|^2_{L^2(\Omega)}.\end{aligned} \tag{5.40}$$

Next, integrating equation (5.39) from 0 to t, and then inserting (5.40) and (5.38) yields

$$\begin{aligned}&\int_0^t \left(\|\dot{w}(s)\|^2_{L^2(\Omega)} + \langle \mathrm{div}\,(w(s)M), \dot{w}(s)\rangle_{L^2(\Omega)}\right) ds + \frac{a}{2}\left(\|\nabla w(t)\|^2_{L^2(\Omega)} - \|\nabla w(0)\|^2_{L^2(\Omega)}\right) \\ &\quad = \int_0^t \left(\|\dot{p}(s) - \dot{\Phi}(s)\|^2_{L^2(\Omega)} + \mathcal{F}_s(w(s), \dot{p}(s) - \dot{\Phi}(s))\right) ds.\end{aligned}$$

Now proceeding as in a) and e), we obtain linear rates under the additional H^1-H^1 regularity of p. Weak solutions to our Fokker–Planck problem are known to preserve the total probability. We conclude this section by proving that this also the case for the semidiscrete solution P.

Lemma 5.3.4. *Let $P = P_N \in H^1(0,T;\mathcal{P}^N_\Omega)$ denote the unique solution of (5.22)–(5.23). Then, for every $t \in [0,T]$, it holds*

$$\int_\Omega P(x,t)\,dx = \int_\Omega P(x,0)\,dx = \sum_{i=1}^N \langle p_0, \psi_i\rangle_{L^2(\Omega)}.$$

Proof. We test (5.22) by $(x \mapsto 1) \in \mathcal{P}^N_\Omega$ and obtain

$$\langle \dot{P}(t), 1\rangle_{L^2(\Omega)} = 0, \quad t \in [0,T].$$

Since P has regularity $H^1(0,T;H^1(\Omega))$, integrating with respect to t yields

$$\int_\Omega P(x,t)\,dx = \int_\Omega P(x,s)\,dx, \quad s,t \in [0,T].$$

□

This concludes the analysis of the semidiscrete Galerkin scheme for FP problem with the state $p = G(u)$. We are now prepared to consider the corresponding semidiscretized cost functional J_N in the next section.

5.4 The semidiscrete cost functional for the Fokker–Planck optimal control problem

In this section, we formulate the semidiscretized cost functionals of $\hat{J}$. Furthermore, we verify the accuracy assumption (C3) from Theorem 5.2.1 using the Galerkin accuracy presented in the previous section. Let $h > 0$ be the spatial mesh size and $N = N(h)$ the dimension of $\mathcal{P}^N_\Omega$.
We introduce the auxiliary semidiscrete problem

$$\min_{u \in U^T_{\text{ad}}} \hat{J}_h(u), \tag{5.41}$$

with $\hat{J}_h(u) := J_N(Y_N(u), u)$ from (5.25) and auxiliary cost functional

$$J_N(y,u) := \frac{\beta}{2}\|y - y^d\|^2_{2,\mathcal{M}} + \frac{\alpha}{2}|y(T) - y^T|^2_{\mathcal{M}} + \frac{\gamma}{2}\|u\|^2_2.$$

The components of the target states $y^d \in H^1(0,T)^N$ and $y^T \in \mathbb{R}^N$ are again defined as the coefficients finite element approximation $\text{Proj}^N_{L^2(\Omega)}$ of p^d and p^T from (5.4), that is,

$$p^d_N(t,x) := \sum_{i=1}^N y^d_i(t)\,\psi_i(x) \text{ and } \quad \|p^d_N - p^d\|_{L^2(\Omega_T)} \le C_\Omega h^2 \|p^d\|_{L^2H^2}, \tag{5.42}$$

$$p^T_N(x) := \sum_{i=1}^N y^T_i\,\psi_i(x) \text{ and } \quad \|p^T_N - p^T\|_{L^2(\Omega)} \le C_\Omega h^2 \|p^T\|_{H^2(\Omega)}. \tag{5.43}$$

A quick computation shows that $\|P - p^d_N\|^2_{L^2(\Omega_T)} = \|y - y^d\|^2_{2,\mathcal{M}}$ and $\|P(T) - p^T_N\|^2_{L^2(\Omega)} = |y(T) - y^T|^2_{\mathcal{M}}$, where $P = G_N(u)$ is the Galerkin approximation of $p = G(u)$ from the previous section. Due to the definition of $\mathcal{M}$ and $\|\cdot\|_{2,\mathcal{M}}$, it holds that

$$\begin{aligned}
\|P - p^d_N\|^2_{L^2(\Omega_T)} &= \int_\Omega \int_0^T \left(\sum_{i=1}^N \big(y_i(t) - y^d_i(t)\big)\psi_i(x) \right)^2 dt\,dx \\
&= \sum_{i=1}^N \sum_{j=1}^N \int_\Omega \int_0^T \psi_i(x)\psi_j(x) \big(y_i(t) - y^d_i(t)\big)^\top \big(y_j(t) - y^d_j(t)\big)\, dt\,dx \\
&= \sum_{i,j=1}^N \int_0^T \big(y_i(t) - y^d_i(t)\big)^\top \mathcal{M}_{ij} \big(y_j(t) - y^d_j(t)\big)\, dt = \|y - y^d\|^2_{2,\mathcal{M}}.
\end{aligned}$$

The second equality is shown analogously. Hence, we also have the following representation of $\hat{J}_h$:

$$\hat{J}_h(u) = \frac{\beta}{2}\|P - p^d_N\|^2_{L^2(\Omega_T)} + \frac{\alpha}{2}\|P(T) - p^T_N\|^2_{L^2(\Omega)} + \frac{\gamma}{2}\|u\|^2_{L^2(0,T)}$$

The next step is to analyze the semidiscrete optimal control problem. For better readability, we consider in the following chapter the special case, where the dimension m of the control is one and $\mathcal{M}$ is the identity matrix. We remark that since $\mathcal{M}$ is a positive definite matrix, the norms $\|\cdot\|_{2,\mathcal{M}}$ and $\|\cdot\|_2$ are equivalent norms on $L^2(0,T)$, and therefore, all the results hold in an analogous way for the mass matrix $\mathcal{M}$ from (5.20).

6 Accuracy estimates for bilinear ODE–constrained optimization problems

Everything should be made as simple as possible, but not simpler.

Albert Einstein, 1879 – 1955

In this chapter, we continue the numerical analysis of the Fokker–Planck optimal control problem from the previous chapter. For this purpose, it is the aim to establish accuracy estimates of a finite element approximation to the optimal control problem

$$\min_{u\in L^2(0,T)} J(y,u), \tag{6.1}$$

with a scalar control $u \in L^2(0,T)$ and $y \in H^1(0,T)^N$ solution to the following linear Cauchy problem with bilinear control mechanism

$$y'(t) = \Big(A + u(t)\,B\Big)\,y(t), \quad t \in [0,T], \qquad y(0) = y_0 \in \mathbb{R}^N. \tag{6.2}$$

We assume that the matrices $A, B \in \mathbb{R}^{N\times N}$ are constant, and we refer the reader to Section 5.3 on the connection of (6.2) to the Fokker–Planck problem. The cost functional $J : H^1(0,T) \times L^2(0,T) \to \mathbb{R}$ is of the quadratic form

$$J(y,u) := \frac{\beta}{2}\int_0^T |y(t) - y_d(t)|^2\,dt + \frac{\gamma}{2}\int_0^T |u(t)|^2\,dt + \frac{\alpha}{2}|y(T) - y_T|^2 \tag{6.3}$$

for $\gamma > 0$ and $\alpha, \beta \geq 0$. We recall Section 5.4, where the objective (6.3) is derived from the Fokker–Planck optimal control problem considered in Chapter 5. The first and last integral quantify the deviation of y and $y(T)$ from a desired state $y_d \in H^2(0,T)^N$ and target configuration $y_T \in \mathbb{R}^N$, respectively. The second integral term represents a L^2–cost of the control and has a strong regularizing effect for solutions to (6.1). The case of additional bilateral box constraints is addressed in the Section 6.8.

It is the aim to define a finite element discretization of problem (6.1)–(6.3) with optimal control $\bar{u}_K$ given on the K points of a uniform grid on $[0,T]$, and to establish a second–order accuracy estimate of the form

$$\|\bar{u} - \bar{u}_K\|_{L^2(0,T)} \leq CK^{-2}, \qquad \text{as } K \to \infty.$$

In the previous chapter, the optimal control problem (6.1) has already been motivated by a semidiscretization of the Fokker–Planck problem with a Galerkin approach. In general, the ordinary differential equation (6.2) arises in many semidiscrete Galerkin schemes to approximate time–dependent partial differential equations with a bilinear control mechanism; see, e.g., [11, 46, 59]. Furthermore, the bilinear structure of the control mechanism in (6.2) appears in, e.g., models of quantum optimal control problems and linearized models of neural networks, c.f. [8–10, 25, 43, 47], and in many application systems [12, 56]. Nevertheless, the numerical approximation of this class of optimal control problems has been less investigated. For this reason, we would like to contribute to this field of research with the development of a numerical analysis framework that is centered, with the aim of large applicability, on first– and second–order optimality conditions. In view of the specific bilinear structure of our problem, we focus on time–discretization of the state and control function by finite elements.

Although there is extensive literature on error estimates of ODE optimization problems, it seems difficult to find results for the problem (6.1)–(6.3) under the given assumptions. This is due to the fact that while most contributions cover a rich variety in the structure of the ODEs, the set of controls, the structure of the cost functional and time discretization schemes, they assume certain coercivity conditions on the optimization problem that are hard to be directly verified in our case. We refer to [29, 39] for a detailed survey of error estimates of non–linear optimal control problems with Runge–Kutta discretizations. The paper [28] examines a wide class of numerical schemes and analyzes the convergence of the first–order optimality system of a non–linear optimal control problem under suitable second–order assumptions on the reduced cost functional. Euler discretizations for an optimal control problem with strong second–order conditions are covered in detail in [2, 27]. The case of linear control mechanisms where this condition may not be fulfilled, as it is the case with a bang–bang control problems, is extensively studied in [3, 53].

However, our approach, based on second–order results from Theorem 1.3.6 and a variational discretization concept [42], is more direct and fits well to the case of bilinear optimal control problems but differs from the classical ones [28, 29, 39]. Furthermore, the use of finite element discretization with the correct finite element spaces allows us to overcome the discrepancy between the discrete optimality system and the discretized optimality system. In general, for traditional time–stepping methods, the discretize–then–optimize (direct) approach and the optimize–then–discretize (indirect) approach do not coincide and therefore complicate the numerical analysis and numerical computations, see [66]. We refer to [41] for a detailed analysis on the difference between the direct and indirect approach for our problem (6.1)–(6.2) with Crank–Nicolson time stepping.

In this chapter, a continuous, piecewise linear finite element discretization is used for the state and adjoint equation. Exploiting the bilinear structure of our problem and due to the concept of variational discretization, we use continuous, piecewise quadratic polynomials to approximate the time–dependent controls. In the framework of finite element discretization, accuracy results for elliptic control problems have been presented in [20, 45]. Furthermore, in [45], a bilinear elliptic control problem is considered, and error estimates of order one for piecewise constant controls and of order 3/2 for continuous, piecewise linear controls are given. Similar results are obtained in [4, 20], where a semilinear elliptic equation is considered with the control term entering linearly.

We remark that our main effort is to analyze problems where the controls are not subject to box constraints. However, in Section 6.8 we discuss extension of our results to the control constrained case.

This chapter is organized as follows. In the next section, we recall basic properties of solutions to (6.2) and analyze the control–to–state map and its Fréchet derivatives. The existence of optimal controls is proven in Section 6.2, and the first–order optimality system is derived. Further, in Section 6.3, we analyze the second–order derivatives of the reduced cost functional and state second–order sufficient conditions for optimality. In Section 6.4, we introduce a finite element discretization scheme of the forward and backward problem. Subsequently, we discuss a variational discretization scheme and derive the discrete optimal control problem in Section 6.5. Section 6.6 is devoted to the convergence of the discrete controls to the corresponding optimal control by using the first– and second–order analysis of the optimization problem and the finite element discretization. With this preparation, we state our main results in Section 6.7, where we derive quadratic error estimates. In Section 6.8, we discuss the case where additional box constraints on the control are present. A numerical algorithm for the computation of optimal controls with our framework is developed in Section 6.9 and results are discussed which support the theoretical findings.

6.1 Analysis of the governing model

In this section, we analyze the Cauchy problem (6.2) for controls $u \in L^2(0,T)$ and fixed initial value $y(0) = y_0 \in \mathbb{R}^N$. By means of the Carathéodory theorem, we say that y is a solution to (6.2) if $y : [0,T] \to \mathbb{R}^N$ is absolutely continuous on $[0,T]$ and satisfies the following equation

$$y(t) = y_0 + \int_0^t \big(A + u(s)B\big)y(s)\,ds, \quad t \in [0,T]. \tag{6.4}$$

Let us recall some notation that is used throughout this chapter. We use the abbreviations for the following norms of only time–dependent Lebesgue spaces

$$\|\cdot\|_2 := \|\cdot\|_{L^2(0,T)}, \quad \|\cdot\|_\infty := \|\cdot\|_{L^\infty(0,T)}, \quad \|\cdot\|_{H^1} := \|\cdot\|_{H^1(0,T)}.$$

For $u, v \in L^2(0,T)$, we denote with $\langle u, v\rangle_2 := \int_0^T u(t)v(t)\,dt$ the canonical scalar product on $L^2(0,T)$. Furthermore, we frequently use the continuous embedding $H^1(0,T) \hookrightarrow C([0,T])$, cf. [1]. We use $|\cdot|$ for the Euclidean norm and matrices $M \in \mathbb{R}^{m\times n}$ are sometimes interpreted as vectors $M \in \mathbb{R}^{mn}$, which gives meaning to $|M|$. In this chapter, all constants C may depend (continuously) on the fixed data $T, |A|, |B|$ and $|y_0|$ from (6.4); any additional dependencies are, if not clear from the context, denoted by subscripts. The following theorem states existence of global solutions of regularity $H^1(0,T)$.

Theorem 6.1.1. *For every $u \in L^2(0,T)$, there exists a unique solution $y \in H^1(0,T)^N$ to (6.4). Furthermore, it holds that*

$$\|y\|_\infty + \|y\|_{H^1} \le C_{T,|y_0|,|A|,|B|,\|u\|_2},$$

i.e., the constant C depends continuously only on the real valued numbers $T, |y_0|, |A|$, $|B|$ and $\|u\|_2$. If $u \in H^1(0,T) \cap C([0,T])$, then $y \in C^1([0,T])^N$ and

$$\|y\|_{H^2} \le C_{\|u\|_{H^1},\|y\|_{H^1}}.$$

Proof. Due to its linear structure, the right–hand side of (6.2) satisfies the Carathéodory condition and is locally Lipschitz continuous in y. Thus, there exists an absolutely continuous, unique solution $y : [0,T] \to \mathbb{R}^N$ with $y' \in L^1(0,T)^N$ which has the representation

$$y(t) = \exp\left(A\,t + B\int_0^t u(s)\,ds\right) y_0, \quad t \in [0,T]. \tag{6.5}$$

This yields the bound for $\|y\|_\infty$. Since y satisfies (6.2) almost everywhere, we obtain the H^1–bound

$$\int_0^T |y'(t)|^2\,dt = \int_0^T \big|(A+u(s)B)y(s)\big|^2\,ds \le C\|y\|_2^2 + C\|y\|_\infty^2\|u\|_2^2.$$

Now let $u \in H^1(0,T)$ be continuous. By standard ODE theory, y from (6.5) is in $C^1([0,T])$ and taking the derivative of (6.2) yields

$$\|y''\|_2 \le |B|\|u'\|_2\|y\|_\infty + \|A+uB\|_\infty\|y'\|_2 \le C\big(1+\|u\|_{H^1}\big)\|y\|_{H^1}.$$

This concludes the proof. □

Notice that, if A and B are skew-symmetric, then the Euclidean norm $|y(t)| = |y_0|$ is preserved for every $t \in [0,T]$, and we refer to [25] for further details on the problem (6.4).
An immediate consequence of Theorem 6.1.1 is the well–definedness of the control–to–state map

$$G : L^2(0,T) \to H^1(0,T)^N, \quad u \mapsto y, \quad y \text{ solution to (6.4) with control } u.$$

Next, we state some properties of the map G.

Lemma 6.1.2. *G is compact in the sense that for every weakly convergent sequence $(u_k) \rightharpoonup u$ in $L^2(0,T)$, the strong convergence $G(u_k) \to G(u)$ in $L^2 \cap L^\infty(0,T)^N$ holds.*

Proof. Since $(u_k) \rightharpoonup u$ in $L^2(0,T)$, the sequence $(\|u_k\|_2)$ is bounded and Theorem 6.1.1 implies the boundedness of $G(u_k)$ in $H^1(0,T)$ uniformly in k. Thus, there exists a limit $y \in H^1(0,T)^N \cap C([0,T])^N$ with $G(u_k) \rightharpoonup y$ in $H^1(0,T)^N$ and due to the compact Sobolev embedding $H(0,T) \Subset C([0,T])$, it holds that $G(u_k) \to y$ uniformly on $[0,T]$ for a subsequence. Lastly, one has to prove that $y = G(u)$. Since solutions to (6.4) are unique and given by the exponential formula (6.5), it is sufficient to show that y fulfills this formula. This is an immediate consequence of the weak $L^2(0,T)$–convergence of the controls, and we obtain

$$y(t) = \lim_{k\to\infty} G(u_k) = \lim_{k\to\infty} \exp\left(A\,t + B\int_0^t u_k(s)\,ds\right) y_0 = \exp\left(A\,t + B\int_0^t u(s)\,ds\right) y_0 = G(u).$$

By the standard argument from Lemma 1.4.3, one can verify a posteriori that this convergence also holds without selecting a subsequence. This completes the proof. □

Lemma 6.1.3. *G is arbitrarily often Fréchet differentiable as a mapping from $L^2(0,T)$ to $H^1(0,T)^N$. Furthermore, it holds that*

a) the first– and second–order derivatives at $u \in L^2(0,T)$ in direction $v \in L^2(0,T)$ are implicitly given by the following systems of ODEs on $[0,T]$:

$$\begin{aligned} \xi := G'(u)v \text{ solves} \quad & \xi'(t) = (A+u(t)B)\xi(t) + v(t)B\,G(u)(t), \quad \xi(0) = 0, \\ \chi := G''(u)(v,v) \text{ solves} \quad & \chi'(t) = (A+u(t)B)\chi(t) + 2v(t)B\,\xi(t), \quad \chi(0) = 0, \end{aligned}$$

b) the solutions $\xi, \chi \in H^1(0,T)^N$ are bounded in $H^1(0,T)^N$ by $C_{\|u\|_{L^2}}\|v\|_2$ and $C_{\|u\|_{L^2}}\|v\|_2^2$, respectively.

c) G is locally Lipschitz continuous, i.e., for every $u_1, u_2 \in L^2(0,T)$ it holds that

$$\|G(u_1) - G(u_2)\|_\infty + \|G(u_1) - G(u_2)\|_{H^1} \le C_{\|u_1\|_2, \|u_2\|_2}\|u_1 - u_2\|_2.$$

Proof. The differentiability follows by a standard argument with the implicit function theorem on Banach spaces. We define the mapping

$$\begin{aligned} &F : H^1(0,T)^N \times L^2(0,T) \to C([0,T])^N \times \mathbb{R}^N, \\ &F(y,u) := \left(t \mapsto \Big(y(t) - y_0 - \int_0^t (A + u(s)B) y(s)\, ds \Big), y(0) - y_0 \right), \end{aligned}$$

where both components are arbitrarily often Fréchet differentiable. Due to Theorem 6.1.1, $F(y,u) = (0,0)$ iff $y = G(u)$, i.e., $F(G(u),u) = (0,0)$ for all $u \in L^2(0,T)$. Due to the linear structure of F in y, it follows by the Carathéodory theorem that the mapping

$$H^1(0,T)^N \ni \xi \mapsto D_y F(y,u)\xi = \left(t \mapsto \Big(\xi(t) - \int_0^t (A + u(s)B)\xi(s)\, ds \Big), \xi(0) \right) \in C([0,T])^N \times \mathbb{R}^N$$

is an isomorphism. Hence, the implicit function theorem states that G is Fréchet differentiable, and differentiating $u \mapsto F(G(u),u)$ with respect to u gives the desired formula for $G'(u)$. An analogous procedure can be done to verify higher-order differentiability and the formula for χ.
Let us mention that one could prove this claim directly by calculating the Fréchet derivative of

$$L^2(0,T) \ni u \mapsto G(u), \quad G(u)(t) = y_0 + \int_0^t \big(A + u(s)B\big) y(s)\, ds.$$

Next, let us show Statement b). Since ξ and χ are solutions to affine linear differential equations, we find $H^1(0,T)$–bounds similarly to the proof of Theorem 6.1.1. Let us define the fundamental solution to the homogeneous part

$$Y(t) := \exp\left(A\,t + B \int_0^t u(s)\, ds \right) \in \mathbb{R}^{n \times n}, \quad t \in [0,T].$$

Since $Y(0)$ is the identity matrix, $\xi(0) = 0$ and $G(u)(t) = Y(t)y_0$, the solution ξ can be stated as

$$\xi(t) = Y(t) \int_0^t Y(s)^{-1} \big(v(s) B\, Y(s) y_0 \big)\, ds, \quad t \in [0,T]. \tag{6.6}$$

Furthermore, we can estimate the exponential matrix

$$|Y(t)| \le \exp\left(\Big| A\,t + B \int_0^t u(s)\, ds \Big| \right) \le C \exp\left(\int_0^t |u(s)|\, ds \right) \le C_{\|u\|_{L^2(0,t)}},$$

and an analogous estimate holds for $Y(t)^{-1}$ for $t \in [0,T]$. Therefore, the $L^\infty(0,T)$–bound of ξ follows from the pointwise estimate

$$|\xi(t)| \le |Y(t)||B| \int_0^t |v(s)||Y^{-1}(s)|Y(s)||y_0|\, ds \le C_{\|u\|_{L^2(0,t)}} \int_0^t |v(s)|\, ds.$$

This yields with $y := G(u)$ the H^1–bound

$$\begin{aligned} \int_0^T |\xi'(t)|^2\, dt &= \int_0^T \big| (A + u(t)B)\xi(s) + v(s)B\, y(s) \big|^2\, dt \\ &\le C\|\xi\|_2^2 + C\|\xi\|_\infty^2 \|u\|_2^2 + \|v\|_2^2 |B| \|y\|_\infty^2. \end{aligned}$$

Consequently, due to the L^∞–bound of y given by Theorem 6.1.1, we obtain $\|\xi\|_{H^1} \le C_{\|u\|_2} \|v\|_2$. We can argue similarly for the second Fréchet derivative χ, which completes the proof of b).
In order to prove the Lipschitz continuity, let $u_1, u_2 \in L^2(0,T)$. Define $y_1 := G(u_1), y_2 := G(u_2)$ and $\delta u := u_1 - u_2$, $\delta y := y_1 - y_2$. A straight forward calculation shows that $\delta y \in H^1(0,T)^N$ solves

$$\delta y' = (A + u_1 B)\delta y + \delta u\, B y_2 \quad \text{a.e. on } [0,T], \qquad \delta y(0) = 0.$$

This is equivalent to the equation for ξ where $\delta y = \xi$ and $\delta u = v$. Thus, we can repeat the estimates from above and obtain the desired Lipschitz estimates

$$\begin{aligned} |y_1(t) - y_2(t)| &\leq C_{\|u_1\|_{L^2(0,t)}} C_{\|u_2\|_{L^2(0,t)}} \int_0^t |u_1(s) - u_2(s)|\, ds, \\ \|y_1 - y_2\|_\infty + \|y_1 - y_2\|_{H^1} &\leq C_{\|u_1\|_2} C_{\|u_2\|_2} \|u_1 - u_2\|_2. \end{aligned} \tag{6.7}$$

This concludes the proof. □

6.2 Analysis of the optimal control problem

In this section, we analyze the optimal control problem (6.1)–(6.3). With the control–to–state map G, we can introduce the reduced cost functional $\hat{J}(u) := J(G(u), u)$ for $u \in L^2(0,T)$, and henceforth, consider the unconstrained minimization problem (P)

$$\min_{u \in L^2(0,T)} \hat{J}(u), \tag{P}$$

where $y_d \in H^2(0,T)$, $y_T \in \mathbb{R}^N$, $\alpha, \beta \geq 0$ and $\gamma > 0$ is given, and

$$\hat{J}(u) = \frac{\beta}{2}\|G(u) - y_d\|_2^2 + \frac{\gamma}{2}\|u\|_2^2 + \frac{\alpha}{2}|G(u)(T) - y_T|^2.$$

In this chapter, the subscript J in a constant C_J means that C_J depends continuously on

$$\|y_d\|_{H^2}, |y_T|, T, \alpha, \beta, \gamma \text{ and } |A|, |B|.$$

Before we perform first– and second–order analysis, let us ensure the well–posedness of the minimization problem (P).

Theorem 6.2.1. *Problem (P) admits at least one solution in the sense that there exists $\bar{u} \in L^2(0,T)$ such that*

$$\hat{J}(\bar{u}) = \inf_{u \in L^2(0,T)} \hat{J}(u) =: \mathcal{I}.$$

Proof. The proof is standard due to the compactness of G and the quadratic form of J. Obviously, $\hat{J}$ is bounded from below, and therefore, there exists a minimizing sequence $(u_k) \subset L^2(0,T)$ with $\lim_{k\to\infty} \hat{J}(u_k) = \mathcal{I} \geq 0$. Since $\gamma > 0$, we obtain the boundedness

$$\limsup_{k\to\infty} \frac{\gamma}{2}\|u_k\|_2^2 \leq \limsup_{k\to\infty} \hat{J}(u_k) = \mathcal{I}.$$

Thus, after selecting a subsequence, there exists a weak limit $\bar{u} \in L^2(0,T)$, and due to Lemma 6.1.2, it holds that $G(u_k) \to G(\bar{u})$ in $L^2(0,T)$. Taking both facts into account, we obtain that $\hat{J}$ is weakly lower semicontinuous, i.e., $\hat{J}(\bar{u}) \leq \liminf_{k\to\infty} \hat{J}(u_k)$. This shows that $\bar{u}$ is indeed a minimizer of $\hat{J}$ due to the estimates $\mathcal{I} \leq \hat{J}(\bar{u}) \leq \liminf_{k\to\infty} \hat{J}(u_k) = \mathcal{I}$. This concludes the proof. □

In order to derive a characterization of solutions to (P), we need to analyze the derivatives of $\hat{J}$. By an application of the chain rule, $\hat{J}$ is as often Fréchet differentiable on $L^2(0,T)$ as G, and therefore, we obtain for $u, v \in L^2(0,T)$ and $y := G(u)$, $\xi := G'(u)v$, $\chi := G''(u)(v,v)$ the following

$$\hat{J}'(u)v = \beta \int_0^T (y(t) - y_d(t)) \cdot \xi(t)\, dt + \alpha(y(T) - y_T) \cdot \xi(T) + \gamma\langle u, v\rangle_2, \tag{6.8}$$

$$\hat{J}''(u)(v,v) = \beta\|\xi\|_2^2 + \beta \int_0^T (y(t) - y_d(t)) \cdot \chi(t)\, dt + \alpha(y(T) - y_T) \cdot \chi(T) + \alpha|\xi(T)|^2 + \gamma\|v\|_2^2. \tag{6.9}$$

Furthermore, since $\hat{J}$ is of class C^∞ in $L^2(0,T)$, its derivatives are locally Lipschitz, and for $u_1, u_2 \in L^2(0,T)$ there exists $C = C_{J,\|u_1\|_2,\|u_1\|_2}$ such that for all $v \in L^2(0,T)$ it holds that

$$|\hat{J}'(u_1)v - \hat{J}'(u_2)v| \leq C\|u_1 - u_2\|_2\|v\|_2, \quad |\hat{J}''(u_1)(v,v) - \hat{J}''(u_2)(v,v)| \leq C\|u_1 - u_2\|_2\|v\|_2^2. \tag{6.10}$$

Next, we need to define the concept of a (local) minimum.

Definition 6.2.2. *Let $\bar{u} \in L^2(0,T)$, and for $\varepsilon > 0$, we recall the following notation for the open L^2–ball*

$$B_\varepsilon(\bar{u}; L^2) = \{u \in L^2(0,T) \mid \|u - \bar{u}\|_2 < \varepsilon\}.$$

We say that $\bar{u}$ is a

i) local minimum of $\hat{J}$ or the local solution of problem (P) if there exists $\varepsilon > 0$ such that for all $u \in B_\varepsilon(\bar{u}; L^2)$ it holds that $\hat{J}(\bar{u}) \leq \hat{J}(u)$, and it is a strict local minimum if $\hat{J}(\bar{u}) < \hat{J}(u)$ for all $u \in B_\varepsilon(\bar{u}; L^2)\backslash\{\bar{u}\}$.

ii) minimum of $\hat{J}$ or the solution of problem (P) if $\hat{J}(\bar{u}) = \mathcal{I}$.

Next, we recall a standard assertion for unconstrained minimization problems with smooth reduced cost functionals. If $\bar{u}$ is a local minimum of $\hat{J}$, then

$$\hat{J}'(\bar{u})v = 0, \quad v \in L^2(0,T).$$

Further, for given $(y,u) \in H^1(0,T)^N \times L^2(0,T)$, we introduce the adjoint variable $q \in H^1(0,T)^N$ defined as the solution to the following Cauchy problem with terminal condition

$$-q'(t) = \beta(y(t) - y_d(t)) + (A + u(t)\,B)^\top q(t), \quad q(T) = \alpha(y(T) - y_T). \tag{6.11}$$

We remark that this differential equation is affine linear in q, and one can show analogously to Section 6.2 that the adjoint problem is well–posed. We also refer to q as backward solution or co–state. Moreover, the following lemma holds.

Lemma 6.2.3. *The control–to–adjoint map*

$$Q : L^2(0,T) \to H^1(0,T)^N \cap C([0,T])^N, \quad u \mapsto q \textit{ solution to } (6.11) \textit{ with } y = G(u)$$

is well-defined, compact (from $L^2(0,T)$ to $L^\infty(0,T)^N$) and locally Lipschitz continuous. Specifically, the following estimates hold for all $u, v \in L^2(0,T)$, $t \in [0,T]$

$$\|Q(u)\|_\infty + \|Q(u)\|_{H^1} \leq C_{\alpha,\beta,y_T,\|y_d\|_2,\|u\|_2}$$
$$|Q(u)(t) - Q(v)(t)| \leq C_{\|u\|_2,\|v\|_2}\|u - v\|_2.$$

Furthermore, if $u \in H^1(0,T) \cap C([0,T])$, there exists a constant C_J (depending on the given quantities of the state equation and the cost functional J) such that the adjoint $Q(u) \in C^1([0,T])$ satisfies

$$\|Q(u)\|_{H^2} \leq C_J C_{\|u\|_{H^1}}$$

Proof. The backward problem for q (for given y) can be transformed into an initial value problem by considering the function $\tilde{q}(t) := q(T-t)$. Thus, existence and uniqueness of absolutely continuous solutions $q = Q(u)$ to (6.11) for $u \in L^2(0,T)$ are again a direct consequence of Carathéodory's existence theorem. Since y and y_d are continuous, and since (6.11) is an affine linear problem, the proof of Lemma 6.1.3 b) and c) can be repeated to obtain the bound in the L^∞–and H^1–norm and the Lipschitz continuity of Q from $L^2(0,T)$ to $L^\infty(0,T)$.

Lastly, we prove compactness. For this purpose, let $u \in L^2(0,T)$ and $u_k \rightharpoonup u$ in $L^2(0,T)$ as $k \to \infty$ be arbitrary but fixed. Due to the boundedness of $\|u_k\|_2$ uniformly in k, the sequence $(\|Q(u_k)\|_{H^1})_{k\in\mathbb{N}}$ is uniformly bounded in k, and there exists a weak H^1 limit $q \in C([0,T])^N$. The compact embedding $H^1(0,T) \Subset C([0,T])$ implies that for a subsequence denoted by $I \subset \mathbb{N}$, it holds that $Q(u_k) \to q$ uniformly in $[0,T]$ and strongly in $L^2(0,T)$ as $I \ni k \to \infty$. It is left to show that $q = Q(u)$. Now let I^* be any countably infinite subset of I and simply observe that for every $k \in \mathbb{N}$

$$Q(u_k)(t) = \alpha\big(G(u_k)(T) - y_T\big) + \int_t^T \big((A + u_k(s)B)^\top Q(u_k)(s) + \beta(G(u_k)(s) - y_d(s))\big)\, ds.$$

The left–hand side converges to $q(t)$ as $I^* \ni k \to \infty$. Due to the compactness of G proven in Lemma 6.1.2 and the strong L^2 convergence of $Q(u_k)$, the right–hand side converges to the desired limit as $I^* \ni k \to \infty$, that is,

$$q(t) = \alpha\big(G(u)(T) - y_T\big) + \int_t^T \big((A + u(s)B)^\top q(s) + \beta(G(u)(s) - y_d(s))\big)\, ds.$$

This proves $q = Q(u)$ by uniqueness. Since the set for the sub–subsequence I^* was arbitrary, the convergence $Q(u_k) \to Q(u)$ holds without selection of a subsequence, see Lemma 1.4.3. This concludes the proof of the compactness. Now let $u \in H^1(0,T)$ and $q := Q(u), y := G(u)$. The H^2–bound of q follows by differentiating (6.11), taking the L^2–norm and applying the H^1–estimate of y from Theorem 6.1.1

$$\begin{aligned}\|q''\|_2 &\le \beta\|y' - y_d'\|_2 + \|A + uB\|_\infty \|q'\|_2 + \|u'\|_2 |B| \|q\|_\infty \\ &\le \beta\big(\|y\|_{H^1} + \|y_d\|_{H^1}\big) + C\big(1 + \|u\|_{H^1}\big)\|q\|_{H^1}.\end{aligned}$$

This concludes the proof. □

In the following lemma, the optimality system is introduced which characterizes local minima.

Lemma 6.2.4. *Let $\bar{u}$ be a local minimum of $\hat{J}$ and define $\bar{y} := G(\bar{u})$, $\bar{q} := Q(\bar{u})$. Then, the triple $(\bar{u}, \bar{y}, \bar{p})$ satisfies for $t \in [0,T]$ the following optimality system*

$$\begin{aligned} \bar{y}'(t) &= (A + \bar{u}(t)B)\bar{y}(t), & \bar{y}(0) &= y_0, \qquad (6.12)\\ -\bar{q}'(t) &= (A + \bar{u}(t)B)^\top \bar{q}(t) + \beta(\bar{y}(t) - y_d(t)), & \bar{q}(T) &= \alpha(\bar{y}(T) - y_T), \qquad (6.13)\\ 0 &= \gamma\bar{u}(t) + \bar{q}(t)^\top B\, \bar{y}(t). & & \qquad (6.14)\end{aligned}$$

Proof. Equations (6.12) and (6.13) hold due to the definition of the G and Q. Now recall the fact that $\hat{J}'(\bar{u})v = 0$ for all $v \in L^2(0,T)$, where $\hat{J}'$ is given by (6.8). Consequently, (6.14) follows directly by testing the equation for the adjoint (6.13) with $\xi = G'(\bar{u})v \in H^1(0,T)$ from Lemma 6.1.3 and then integrating by parts as this implies

$$\beta\langle \bar{y} - y_d, \xi\rangle_2 = \langle \bar{q}, vB\bar{y}\rangle_2 + \bar{q}(T)\xi(T),$$

and thus, $\hat{J}'(\bar{u})v = \gamma\langle \bar{u}, v\rangle_2 + \langle \bar{q}^\top B\, \bar{y}, v\rangle_2$. From this variational formulation, we obtain equation (6.14) a.e. on $[0,T]$ by the fundamental lemma of calculus of variations. □

Due to the optimality system (6.12)–(6.14), we deduce higher regularity of a local minimum of $\hat{J}$ as follows.

Corollary 6.2.5. *Let $\bar{u}$ be a local minimum of $\hat{J}$. Then, the functions $\bar{u}$, $\bar{y} := G(\bar{u})$ and $\bar{q} := Q(\bar{u})$ have the higher regularity $C^2([0,T])$, and the following implicit equation holds for $t \in [0,T]$*

$$\bar{u}(t) = -\frac{1}{\gamma}\bar{q}(t)^\top B\, \bar{y}(t). \qquad (6.15)$$

Furthermore, we have the following estimate

$$\|\bar u\|_{H^1} \le C_\gamma \|\bar q\|_{H^1} \|\bar y\|_{H^1}.$$

Proof. The first claim follows by a bootstrap argument. Since $\bar u \in L^2(0,T)$, the functions $G(\bar u)$ and $Q(\bar u)$ are at least in $H^1(0,T)^N$. Now, the assumption $\gamma > 0$ together with equation (6.14) and the fundamental lemma of the calculus of variations yield $\bar u = -\frac{1}{\gamma}\bar q^\top B\,\bar y$ almost everywhere on $[0,T]$. This implies $\bar u \in W^{1,1}(0,T) \cap C([0,T])$ after (possibly) changing $\bar u$ on a set of measure zero; note that this procedure has no effect on $G(\bar u), Q(\bar u)$ or $\hat J(\bar u)$. Therefore, all coefficients of the affine linear system (6.12)–(6.13) are continuous, and thus, by standard ODE theory, we obtain that $\bar y, \bar q \in C^1([0,T])^N$. Once again, we make use of the above representation of $\bar u$ which yields C^1–regularity. Since the target state y_d has H^2–regularity, we can repeat this procedure one more time to conclude $\bar y, \bar q \in C^2([0,T])^N$, and therefore, we deduce that $\bar u \in C^2([0,T])$.
The bound on $\bar u$ in the H^1–norm is obtained by taking the derivative of its implicit representation

$$\gamma\|\bar u'\|_2 \le \big(\|q'^\top y\|_2 + \|q^\top y'\|_2\big)|B| \le \big(\|q'\|_2\|y\|_\infty + \|q\|_\infty\|y'\|_2\big)|B|. \tag{6.16}$$

This concludes the proof. □

Notice that Corollary 6.2.5 allows us to differentiate $\bar u$ with respect to t, and hence we obtain with (6.12) and (6.13)

$$\begin{aligned}
-\gamma\frac{d}{dt}\bar u(t) &= \bar q'(t)^\top B\,\bar y(t) + \bar q(t)^\top B\,\bar y'(t)\\
&= -\Big((A+\bar u(t)B)^\top \bar q(t) + \beta(\bar y(t) - y_d(t))\Big)^\top B\,\bar y(t) + \bar q(t)^\top B\,\big(A + \bar u(t)B\big)\bar y(t)\\
&= -\bar q(t)^\top (A+\bar u(t)B)B\,\bar y(t) - \beta(\bar y(t)-y_d(t))^\top B\,\bar y(t) + \bar q(t)^\top B\,\big(A+\bar u(t)B\big)\bar y(t).
\end{aligned}$$

which implies, if $AB = BA$, the following representation

$$\bar u'(t) = \frac{\beta}{\gamma}(\bar y(t) - y_d(t))^\top B\,\bar y(t), \quad t \in [0,T].$$

The right–hand side is independent of the adjoint, and therefore, it can be interpreted as an (infinitesimal) feedback–mechanism for an optimal control. Similarly, at the final time we have

$$\bar u(T) = -\frac{\alpha}{\gamma}(\bar y(T) - y_T)^\top B\,\bar y(T),$$

and if $\beta = 0$, then every optimal control is constant (with this value). Another consequence of the implicit equation is a (standard) uniqueness result for optimal controls if γ is sufficiently large compared to the data A, B, T, y_d, y_T. Furthermore, we have proven the following integro–differential equation for minimizers.

Corollary 6.2.6. *Consider the set* $\mathcal U_0 := \{u \in C^1([0,T]) \mid u \text{ solves } (6.17)\text{–}(6.18)\}$, *where*

$$u'(t) = \frac{\beta}{\gamma}\left(e^{At}\exp\left(B\int_0^t u(s)\,ds\right)y_0 - y_d(t)\right)^\top Be^{At}\exp\left(B\int_0^t u(s)\,ds\right)y_0, \tag{6.17}$$

$$u(T) = -\frac{\alpha}{\gamma}\left(e^{AT}\exp\left(B\int_0^T u(s)\,ds\right)y_0 - y_T\right)^\top Be^{AT}\exp\left(B\int_0^T u(s)\,ds\right)y_0. \tag{6.18}$$

If A, B *commute, then every local minimizer* $\bar u$ *lies in* $\mathcal U_0$ *and* $\mathcal U_0 = \ker(\hat J')$.

Corollary 6.2.5 can be quite useful if one considers the case $\alpha = 0$ or $\beta = 0$. If $\beta = 0$ and $\alpha \neq 0$, we have $u' = 0$ on $[0,T]$ and hence the constant value of the control can be directly computed from (6.18), that is, find the unique $x = u(T) \in \mathbb{R}$ such that

$$x = -\frac{\alpha}{\gamma}\left(e^{AT}e^{BTx}y_0 - y_T\right)^\top Be^{AT}e^{BTx}y_0.$$

Furthermore, we have proven uniqueness of the minimizer in that case, since $\hat{J}$ cannot have any other critical points.

In order to obtain error estimates, coercivity of the quadratic form $v \mapsto \hat{J}''(\cdot)(v,v)$ in a neighborhood of a strict local minimum $\bar{u}$ is necessary in the sense that there exists $\varepsilon, \Lambda > 0$ such that $\hat{J}''(u)(v,v) \geq \Lambda\|v\|_2^2$ for all $v \in L^2(0,T)$ and $u \in B_\varepsilon(\bar{u}; L^2)$. Due to the non–linearity of the control–to–state map, it is in general difficult to transfer the convexity in u of our cost functional J to the desired second–order conditions of the reduced cost functional $\hat{J}$. However, in our specific case, we can exploit the bilinear structure of the state equation and make use of results of modern theory of second–order analysis for non–convex minimization problems as given in [21].

6.3 Second–order analysis

It is the aim of this section to apply Theorem 1.3.6 to our optimal control problem. For that purpose, we need to verify the following properties of $\hat{J}$.

Lemma 6.3.1. *For all $(u_k), (v_k) \subset L^2(0,T)$ with $u_k \to u$ and $v_k \rightharpoonup v$ in $L^2(0,T)$, it holds that*

a) $\hat{J}'(u)v = \lim_{k\to\infty} \hat{J}'(u_k)v_k$;

b) $\hat{J}''(u)(v,v) \leq \liminf_{k\to\infty} \hat{J}''(u_k)(v_k,v_k)$;

c) if $v = 0$, then $\gamma \liminf_{k\to\infty} \|v_k\|_2^2 \leq \liminf_{k\to\infty} \hat{J}''(u)(v_k,v_k)$.

We remark that the second derivative of $\hat{J}$ at u is a bilinear mapping from $L^2(0,T) \times L^2(0,T)$ to $\mathbb{R}$.

Proof. Recall that for $u,v \in L^2(0,T)$ and $y := G(u)$, $\xi := G'(u)v$, $\chi := G''(u)(v,v)$,

$$\hat{J}'(u)v = \beta\int_0^T (y(t) - y_d(t)) \cdot \xi(t)\,dt + \alpha(y(T) - y_T) \cdot \xi(T) + \gamma\langle u, v\rangle_2, \tag{6.19}$$

$$\hat{J}''(u)(v,v) = \beta\|\xi\|_2^2 + \beta\int_0^T (y(t) - y_d(t)) \cdot \chi(t)\,dt + \alpha(y(T) - y_T) \cdot \chi(T) + \alpha|\xi(T)|^2 + \gamma\|v\|_2^2. \tag{6.20}$$

First, notice that the Lipschitz continuity (6.10) implies – if the limits exist –

$$\lim_{k\to\infty} \hat{J}'(u_k)v_k = \lim_{k\to\infty}\left(\hat{J}'(u_k) - \hat{J}'(u)\right)v_k + \lim_{k\to\infty} \hat{J}'(u)v_k = \lim_{k\to\infty} \hat{J}'(u)v_k. \tag{6.21}$$

Similarly, we have $\liminf_{k\to\infty} \hat{J}''(u_k)(v_k,v_k) = \liminf_{k\to\infty} \hat{J}''(u)(v_k,v_k)$. Next, let us define $\xi_k := G'(u)v_k$, $\chi_k := G''(u)(v_k,v_k)$ for $k \in \mathbb{N}$. Analogously to the compactness of G, one can show the convergences $\xi_k \to \xi$, $\chi_k \to \chi$ in $L^2(0,T)$ and uniformly on $[0,T]$. This is obviously sufficient for the convergences $\hat{J}'(u)v_k \to \hat{J}'(u)v$ and $\liminf_{k\to\infty} \hat{J}''(u)(v_k,v_k) \geq \hat{J}''(u)(v,v)$; for the latter, recall the weak lower semicontinuity of the L^2–norm. This fact together with equation (6.21) proves the assertions a) and b) of Lemma 6.3.1.

In order to prove c), let $v_k \rightharpoonup 0$ in $L^2(0,T)$. The above convergence results imply $\xi_k, \chi_k \to 0$ in $L^2(0,T)$ and uniformly on $[0,T]$, and thus,

$$\beta\|\xi_k\|_2^2 + \beta\int_0^T (y(t) - y_d(t)) \cdot \chi_k(t)\,dt + \alpha(y(T) - y_T) \cdot \chi_k(T) + \alpha|\xi_k(T)|^2 \to 0, \quad k \to \infty.$$

This immediately implies $\gamma \liminf_{k\to\infty} \|v_k\|_2^2 = \liminf_{k\to\infty} \hat{J}''(u)(v_k, v_k)$ which concludes the proof of Lemma 6.3.1. □

The following theorem states that, in our setting, the positive definiteness $\hat{J}''(\bar{u})(v,v) > 0$ is a sufficient second–order condition for $\bar{u}$. It is a consequence of Lemma 6.3.1 and Theorem 1.3.6; notice that in our case there are no box constraints on the controls.

Theorem 6.3.2. *Let $\bar{u} \in L^2(0,T)$ fulfill the second–order condition $\hat{J}''(\bar{u})(v,v) > 0$ for all $v \in L^2(0,T)\backslash\{0\}$ and $\hat{J}'(\bar{u}) = 0$ on $L^2(0,T)$. Then, $\bar{u}$ is a strict local minimum of $\hat{J}$, and there exist ε, δ, $\Lambda > 0$ such that for all $u \in B_\varepsilon(\bar{u}; L^2)$ the following holds:*

i) the quadratic growth condition $\hat{J}(\bar{u}) + \frac{\delta}{2}\|u - \bar{u}\|_2^2 \leq \hat{J}(u)$;

ii) if $\hat{J}'(u)v = 0$ for all $v \in L^2(0,T)$, then $u = \bar{u}$;

iii) local coercivity of the second derivative $\hat{J}''(u)(v,v) \geq \frac{\Lambda}{2}\|v\|_2^2$ for all $v \in L^2(0,T)$.

This section concludes the analysis of the continuous optimal control problem. Next, we focus on its discretized counterpart.

6.4 Finite element discretization of the forward and backward problems

In this section, we introduce a linear finite element discretization scheme for the forward and backward problem. For this purpose, we define a uniform grid on $[0,T]$ with $K \in \mathbb{N}$ subintervals of length $\Delta t := T/K$ and the grid points

$$t_i := i\Delta t \quad \text{for } i = 0, \dots, K.$$

On this grid, we introduce the following finite–dimensional spaces of polynomials

$$\begin{aligned}
\mathcal{P}_K^0 &:= \{\psi : [0,T[\to \mathbb{R} \mid \psi \text{ is constant on } [t_i, t_{i+1}[\,, i = 0, \dots, K-1\} \subset L^2(0,T), \\
\mathcal{P}_K^1 &:= \{\psi \in C([0,T]) \mid \psi \text{ is linear on } [t_i, t_{i+1}],\ i = 0, \dots, K-1\} \subset H^1(0,T), \\
\mathcal{P}_K^2 &:= \{\psi\phi \mid \psi, \phi \in \mathcal{P}_K^1\}.
\end{aligned}$$

We refer to $\mathcal{P}_K^1$ as the space of continuous, piecewise linear polynomials and to $\mathcal{P}_K^2$ as the space of continuous, piecewise quadratic polynomials. Notice that the set of hat functions $\{\psi_i \mid i = 0, \dots, N\}$ at grid points t_i forms a basis of $\mathcal{P}_K^1$, where for $i = 1, \dots, K-1$, we have

$$\psi_i(t) := \begin{cases} (t - t_{i-1})/\Delta t, & t \in [t_{i-1}, t_i[, \\ (t_{i+1} - t)/\Delta t, & t \in [t_i, t_{i+1}[, \\ 0, & \text{else}. \end{cases} \tag{6.22}$$

Further, at the end points, we have

$$\psi_0(t) := \begin{cases} (t_1 - t)/\Delta t, & t \in [0, t_1[\,, \\ 0, & \text{else}, \end{cases} \quad \text{and} \quad \psi_K(t) := \begin{cases} (t - t_{K-1})/\Delta t, & t \in [t_{K-1}, T[\\ 0, & \text{else}. \end{cases} \tag{6.23}$$

The maps $\mathrm{Proj}_K^1 : H^1(0,T) \to \mathcal{P}_K^1$ and $\mathrm{Proj}_K^2 : H^1(0,T) \to \mathcal{P}_K^2$ denote the L^2–orthogonal projection from (continuous) $H^1(0,T)$ functions to continuous and piecewise quadratic functions. There exists $C = C_T$ such that

$$\|f - \mathrm{Proj}_K^1(f)\|_2 \leq C\|f\|_{H^2}K^{-2}, \quad \|f - \mathrm{Proj}_K^1(f)\|_{H^1} \leq C\|f\|_{H^2}K^{-1}, \qquad f \in H^2(0,T); \tag{6.24}$$

$$\|f - \mathrm{Proj}_K^1(f)\|_2 \leq C\|f\|_{H^1}K^{-1}, \quad \lim_{K\to\infty} \|f - \mathrm{Proj}_K^1(f)\|_{H^1} = 0, \qquad f \in H^1(0,T); \tag{6.25}$$

for a proof we refer to [33, Proposition 1.5]. The same convergence rates hold for Proj_K^2.

Next, we introduce the discrete control–to–state map $G_K : L^2(0,T) \to (\mathcal{P}_K^1)^N$ and the discrete control–to–adjoint map $Q_K : L^2(0,T) \to (\mathcal{P}_K^1)^N$: For given $u \in L^2(0,T)$, find $y_K, q_K \in (\mathcal{P}_K^1)^N$ such that, for all $i = 0, \dots, K$, it holds

$$\langle y_K', \psi_i\rangle_2 = \langle (A+uB)y_K, \psi_i\rangle_2, \tag{6.26}$$

$$-\langle q_K', \psi_i\rangle_2 = \langle (A+uB)^\top q_K + \beta(y_K - y_{d,K}), \psi_i\rangle_2, \tag{6.27}$$

with $y_K(0) = y_0$, $q_K(T) = \alpha(y_K(T) - y_T)$ and $y_{d,K} := \mathrm{Proj}_K^1(y_d)$. This setting (formally) defines the mappings

$$G_K, Q_K : L^2(0,T) \to (\mathcal{P}_K^1)^N, \quad G_K(u) := y_K,\ Q_K(u) := q_K.$$

Notice that the well–definedness of $u \mapsto G_K(u), Q_K(u)$ follows if K is sufficiently large compared to $|A|, |B|, \|u\|_2$. This can be seen by performing the integrations in (7.6)–(7.7). Thus, these variational equations become implicit (algebraic) equations for the grid values $y_K(t_i), q_K(t_i)$ for $i = 0, \dots, K$, when the values $y_K(t_j), q_K(t_k)$ for $j < i,\ , k > i$ are known. For $y^i := y_K(t_i)$ we obtain for $i = 1, \dots K-1$ the left–hand side of (7.6)

$$\langle y_K', \psi_i\rangle_2 = \frac{1}{2}(y^{i+1} - y^{i-1}),$$

and for the right–hand side

$$\begin{aligned}\langle (A+uB)y_K, \psi_i\rangle_2 =& \frac{\Delta t}{6} A(y^{i-1} + 4y^i + y^{i+1}) + \int_{t_{i-1}}^{t_i} u(t) \left(\frac{t - t_{i-1}}{\Delta t} - \frac{(t-t_{i-1})^2}{\Delta t^2}\right) dt\, By^{i-1} \\ &+ \langle u, \psi_i^2\rangle_2 y^i + \int_{t_i}^{t_{i+1}} u(t) \left(\frac{t_{i+1} - t}{\Delta t} - \frac{(t_{i+1} - t)^2}{\Delta t^2}\right) dt\, By^{i+1}\end{aligned}$$

with obvious modifications for $i = 0$ and $i = K$. Analogous equations can be obtained for (7.7). We can uniquely solve the equations for $y_K(t_i)$ and $q_K(t_i)$ if the integral of u over $[t_i, t_{i+1}]$ is sufficiently small. In the special case of piecewise constant controls, (7.6) becomes for $i = 1, \dots K-1$

$$y^{i+1} - y^{i-1} = \Delta t \left(\frac{1}{3}(A + u^{i-1}B)y^{i-1} + \frac{4}{3}(A + u^i By^i) + \frac{1}{3}(A + u^{i+1}B)y^{i+1}\right),$$

which in the framework of linear multistep methods is known as the Milne–method; see Lemma 6.9.1.

It is sufficient for our analysis to define G_K, Q_K on some set $\{u \in L^2(0,T) \mid \|u\|_2 \le R\}$; see below (6.31) for the definition of R. This yields the following well–definedness result:

Lemma 6.4.1. *There exists $K_0 \in \mathbb{N}$ such that for all $K \ge K_0$ the mappings G_K, Q_K are well–defined for controls from the set $\{u \in L^2(0,T) \mid \|u\|_2 \le R\}$.*

In the following we assume $K_0 = 1$ for shorter notations. If u is a piecewise polynomial, we can carry out all integrations in the variational formulation (7.6)–(7.7). As demonstrated in Section 6.9 for $u \in \mathcal{P}_K^2$, we obtain an implicit, linear multistep scheme [18, Chapter 4] for the grid values of y_K and q_K. We refer to [36, Chapter 3.3] on the connection between finite elements and finite difference schemes in general.

Next, we examine its accuracy. First, notice that our first–order problem and its finite element formulation differs from the classical case for elliptic equations that satisfy the usual coercivity condition. In this setting, second–order accuracy holds for L^2–controls u by an application of (6.24) and Céa's lemma, cf. [20] and [61, Chapter 14]. In our framework, we have second–order accuracy results for G_K and Q_K at least for smooth controls $u \in C^2([0,T])$

$$\|G_K(u) - G(u)\|_2 + \|Q_K(u) - Q(u)\|_2 \le C_{\|\bar{u}''\|_\infty} K^{-2}. \tag{6.28}$$

This is proven in Lemma 6.9.1 from Section 6.9 by using (6.24) and with techniques from linear multistep methods. For piecewise polynomials $u_K \in \mathcal{P}_K^2$, however, the corresponding exact solutions $G(u_K)$ and $Q(u_K)$ only belong to $H^2(0,T)$ and it seems difficult to prove or find second–order accuracy results in that case. In Lemma 6.9.2 from Section 6.9, we are able to show that at least first–order accuracy holds for $u_K \in \mathcal{P}_K^2$, that is,

$$\|G_K(u_K) - G(u_K)\|_2 + \|Q_K(u_K) - Q(u_K)\|_2 \leq C_{\|u_K\|_{H^1}} K^{-\sigma} \tag{6.29}$$

with $\sigma = 1$. In view of (6.24), (6.28) and the numerical evidence presented in Section 6.9, it seems reasonable to assume that (6.29) with $\sigma = 2$ can be shown with a more sophisticated approach.

The Fréchet derivative of $G_K : L^2(0,T) \to \mathbb{R}$ can be obtained analogously to the continuous case. Hence, for $u, v \in L^2(0,T)$, we have that $\xi_K := G'_K(u)v$ satisfies the following problem

$$\langle \xi'_K, \psi_i \rangle_2 = \langle (A + uB)\xi_K, \psi_i \rangle_2 + \langle vB\, y_K, \psi_i \rangle_2, \quad \xi_K(0) = 0, \qquad i = 0, \ldots, K. \tag{6.30}$$

6.5 A variational discretization of the optimal control problem

Central to the idea of a variational discretization scheme is the following observation. If the discrete state and adjoint belong to a finite element space, then the minimizer of the discrete cost functional also belongs to a certain finite element space. In other words, we only have to discretize the control–to–state and control–to–adjoint map but not the space $L^2(0,T)$ over which we minimize, and we still obtain that local minimizers of this (semidiscrete) problem belong to a finite–dimensional functional space. In our case, this approach works due to the bilinear structure of the problem, and the key idea is taking advantage of the implicit formula of local minimizers.

To start, we introduce the discretized reduced cost functional $\hat{J}_K$. For this purpose, we replace the continuous quantities $G(u)$ and y_d with their discretized, finite–dimensional counterpart. This gives rise to the functional

$$\begin{aligned} &\hat{J}_K : L^2(0,T) \to \mathbb{R}, \\ &\hat{J}_K(u) := \frac{\beta}{2} \int_0^T |y_K(t) - y_{d,K}(t)|^2 \, dt + \frac{\gamma}{2} \|u\|_2^2 + \frac{\alpha}{2} |y_K(T) - y_T|^2, \end{aligned} \tag{6.31}$$

where $y_K = G_K(u)$ and $y_{d,K} = \mathrm{Proj}_K^1(y_d)$ are $(\mathcal{P}_K^1)^N$ functions. If we look for minimizers $\bar{u}_K$ of $\hat{J}_K$, we obtain an upper bound $R > 0$ for $\|\bar{u}_K\|_2$ by calculating $\hat{J}_K(0)$, since all the terms in (6.31) are non–negative. The bound R depends on $|A|, |B|, \|y_d\|_{H^1}, |y_T|, \alpha, \beta, \gamma$ but not on K. Analogously to Theorem 6.2.1, one can prove existence of minimizers.

Lemma 6.5.1. *The semidiscrete optimal control problem (we call it in this instance semidiscrete because we did discretize the state space but not the control space $L^2(0,T)$)*

$$\min_{u \in L^2(0,T)} \hat{J}_K(u) \tag{P_K}$$

admits a global minimizer $\bar{u}_K$, i.e. $\hat{J}_K(\bar{u}_K) = \inf_{u \in L^2(0,T)} \hat{J}_K(u)$.

The next step is to analyze the functional structure of solutions to (P_K). For this purpose, the following theorem states that for our bilinear optimal control problem with piecewise linear discretization in the forward and backward problem, the space of piecewise quadratic polynomials is the natural discrete optimization space. In fact, we conclude that the discrete optimality system and the discretized optimality system do coincide with our choices of finite element spaces. Notice that the concept of local minima of $\hat{J}_K$ (over $L^2(0,T)$) is defined analogously to Definition 6.2.2.

Theorem 6.5.2. *Let G_K, Q_K be given by (7.6)–(7.7). Then the reduced cost functional $\hat{J}_K$ is Fréchet differentiable from $L^2(0,T)$ to $\mathbb{R}$. Further, let $\bar{u}_K \in L^2(0,T)$ be a local minimum of $\hat{J}_K$, and define $\bar{y}_K := G_K(\bar{u}_K)$, $\bar{q}_K := Q_K(\bar{u}_K)$. Then the following holds:*

a) For every $u \in L^2(0,T)$, we have

$$\hat{J}_K'(u)v = \gamma\langle \bar{u}, v\rangle_2 + \langle q_K^\top B\, y_K, v\rangle_2, \qquad v \in L^2(0,T), \tag{6.32}$$

where $y_K := G_K(u)$, $q_K := Q_K(u)$.

b) For every $u \in L^2(0,T)$, the first–order derivative $\hat{J}_K'(u) \in L^2(0,T)'$ can be identified as a $L^2(0,T)$ function that has the pointwise formula

$$\hat{J}_K'(u)(t) = \gamma u(t) + q_K(t)^\top B\, y_K(t), \quad \textit{f.a.e. } t \in [0,T],$$

where $y_K := G_K(u)$, $q_K := Q_K(u)$.

c) $\bar{u}_K$ belongs to the finite–dimensional space $\mathcal{P}_K^2$ and has the representation

$$\bar{u}_K(t) = -\frac{1}{\gamma}\bar{q}_K(t)^\top B\, \bar{y}_K(t), \quad t \in [0,T]. \tag{6.33}$$

Specifically, the minimization problems of $\hat{J}_K$ over the finite–dimensional space $\mathcal{P}_K^2$ and over $L^2(0,T)$ are equivalent in the following sense:

$$\bar{u} \in L^2(0,T) \textit{ is a local minimum of } \hat{J}_K|_{L^2(0,T)} \iff \bar{u} \in \mathcal{P}_K^2 \textit{ is a local minimum of } \hat{J}_K|_{\mathcal{P}_K^2}.$$

Proof. The Fréchet differentiability of G_K implies the differentiability of $\hat{J}_K$ on $L^2(0,T)$ by an application of the chain rule. For $u, v \in L^2(0,T)$, we define $\xi_K := G_K'(u)v \in (\mathcal{P}_K^1)^N$ and $y_K := G_K(u)$, and we obtain the following formula

$$\hat{J}_K'(u)v = \beta\int_0^T (y_K(t) - y_{d,K}(t))\xi_K(t)\,dt + \alpha(y_K(T) - y_T)\xi_K(T) + \gamma\langle u, v\rangle_2.$$

Next, we derive the representation of $J_K'(u)$ via the discrete adjoint $q_K := Q_K(u)$; similarly to the continuous case. To this end we need to show that the following holds

$$\beta\langle y_K - y_{d,K}, \xi_K\rangle_2 + \alpha(y_K(T) - y_T)\xi_K(T) = \langle q_K^\top B\, y_K, v\rangle_2 \qquad v \in L^2(0,T). \tag{6.34}$$

First, recall the equations (7.7) and (6.30) that determine $\bar{q}_K$ and ξ_K. Due to the linearity in ψ_i, and since $\{\psi_i\}$ form a basis of $\mathcal{P}_K^1$, we obtain for all $\phi \in (\mathcal{P}_K^1)^N$ the following

$$-\langle \bar{q}_K', \phi\rangle_2 = \langle (A + uB)^\top \bar{q}_K + \beta(\bar{y}_K - y_{d,K}), \phi\rangle_2, \quad \bar{q}_K(T) = \alpha(\bar{y}_K(T) - y_T), \tag{6.35}$$

$$\langle \xi_K', \phi\rangle_2 = \langle (A + uB)\xi_K, \phi\rangle_2 + \langle vB\,\bar{y}_K, \phi\rangle_2, \quad \xi_K(0) = 0. \tag{6.36}$$

Consequently, we obtain (6.34) if we test (6.35) with $\phi = \xi_K \in (\mathcal{P}_K^1)^N$, partially integrate and test (6.36) with $\phi = \bar{q}_K \in (\mathcal{P}_K^1)^N$. This yields the variational representation

$$\hat{J}_K'(u)v = \gamma\langle u, v\rangle_2 + \langle \bar{q}_K^\top B\, \bar{y}_K, v\rangle_2, \quad \text{for all } u, v \in L^2(0,T).$$

We apply the Riesz representation theorem and the fundamental lemma of the calculus of variation to find that $v \mapsto \hat{J}_K'(u)v$ can be represented as $L^2(0,T)$–function with pointwise formula

$$\hat{J}_K'(u)(t) = \gamma u(t) + q_K(t)^\top B\, y_K(t), \quad \text{f.a.e. } t \in [0,T].$$

This concludes the proof of part b).
Now let $\bar{u}_K \in L^2(0,T)$ be a local minimizer and let $\bar{y}_K := G_K(\bar{u}_K), \bar{q}_K = Q_K(\bar{u}_K)$. Therefore, it must hold $\hat{J}_K'(\bar{u}_K)(t) = 0$, i.e.

$$\bar{u}_K(t) = -\frac{1}{\gamma}\bar{q}_K(t)^\top B\,\bar{y}_K(t), \quad \text{f.a.e. } t \in [0,T].$$

Apparently, the right–hand side is the product of piecwise linear polynomials. Therefore, we obtain formula (6.33) and $\bar{u}_K \in \mathcal{P}_K^2$ after (possibly) modifying $\bar{u}_K$ on a set of measure zero. The last claim follows from the inclusion $\mathcal{P}_K^2 \subset L^2(0,T)$. □

Our main steps for deriving error estimates are presented in the next section.

6.6 Convergence of optimal controls of the discretized problem

First, we present preliminary results for our error estimates by studying the convergences as K tends to infinity of the discrete maps G_K, Q_K, $\hat{J}_K'$ and the local discrete minima $\bar{u}_K$. We start with the discrete control–to–state and discrete control–to–adjoint maps. Recall that local minima of $\hat{J}$ and $\hat{J}_K$ are elements of $C^2([0,T])$ and $\mathcal{P}_K^2$, respectively.

Lemma 6.6.1. *The following holds for local minima $\bar{u}$ and $\bar{u}_K$ of $\hat{J}$ and $\hat{J}_K$, respectively:*

a) there exists $C^{(1)} = C_{J,\|\bar{u}\|_{H^1},\|\bar{u}_K\|_{H^1}} > 0$ such that for all $K \in \mathbb{N}$

$$\|G(\bar{u}) - G_K(\bar{u}_K)\|_2 + \|Q(\bar{u}) - Q_K(\bar{u}_K)\|_2 \le C^{(1)}\left(K^{-2} + \|\bar{u} - \bar{u}_K\|_2\right);$$

b) let $(u_K)_{K\in\mathbb{N}}$ with $u_K \in \mathcal{P}_K^0 \cup \mathcal{P}_K^2$ and $u_K \rightharpoonup \bar{u}$ in $L^2(0,T)$ as $K \to \infty$. Then it holds that

$$\hat{J}(\bar{u}) \le \liminf_{K\to\infty} \hat{J}_K(u_K),$$

and we have strong convergences of the sequences

$$G_K(u_K) \to G(u), \quad Q_K(u_K) \to Q(u) \text{ in } L^\infty(0,T) \quad \text{as } K \to \infty.$$

Proof. Assertion a) follows from the triangle inequality by inserting $G(\bar{u}_K)$ and $Q(\bar{u}_K)$, respectively, and then exploiting the Lipschitz continuity of G and Q and the convergence rate for G_K and Q_K given in (6.29).
The strong convergences from part b) follow, after inserting the terms $G(u_K), Q(u_K)$ and making use of the triangle inequality, from the compactness of G and Q; see Lemma 6.1.2 and Lemma 6.2.3. These results, together with $y_{d,K} \to y_d$ in $L^\infty(0,T)$ and $\|\bar{u}\|_2^2 \le \liminf_{K\to\infty}\|u_K\|_2^2$, imply with (6.31) the desired estimate $\hat{J}(\bar{u}) \le \liminf_{K\to\infty}\hat{J}_K(u_K)$ and the proof is complete. □

Next, we provide convergence results for the reduced gradient.

Lemma 6.6.2. *Let $\bar{u}$ and $\bar{u}_K$ be some local minima of $\hat{J}$ and $\hat{J}_K$. Then, there exists $C^{(2)} = C_{J,\|\bar{u}_K\|_{H^1},\|\bar{u}\|_{H^1}} > 0$ such that for all $v \in L^2(0,T)$, it holds for $u \in \{\bar{u}_K, \bar{u}\}$ that*

$$\left|\left(\hat{J}_K'(\bar{u}_K) - \hat{J}'(u)\right)v\right| \le C^{(2)}\left(K^{-2} + \|\bar{u}_K - u\|_2\right)\|v\|_2.$$

Proof. First, recall the reduced gradients for $u \in \{\bar{u}_K, \bar{u}\}$

$$\hat{J}_K'(\bar{u}_K)v = \gamma\langle\bar{u}_K, v\rangle_2 + \langle Q_K(\bar{u}_K)^\top B\,G_K(\bar{u}_K), v\rangle_2,$$
$$\hat{J}'(u)v = \gamma\langle u, v\rangle_2 + \langle Q(u)^\top B\,G(u), v\rangle_2.$$

Therefore, by the Cauchy-Schwarz inequality, it follows that

$$\left|\left(\hat{J}'_K(\bar{u}_K) - \hat{J}'(u)\right) v\right| \leq \gamma \|\bar{u}_K - u\|_2 \|v\|_2 + \|Q_K(\bar{u}_K)^\top B\, G_K(\bar{u}_K) - Q(u)^\top B\, G(u)\|_2 \|v\|_2.$$

Due to Lemma 6.6.1 and using triangle inequalities, we can estimate the last term against $C^{(1)}(K^{-2} + \|\bar{u}_K - u\|_2)$ which concludes the proof. □

The following theorem establishes the relationship between the discrete minimizer $\bar{u}_K \in \mathcal{P}^2_K$ of problem (P_K) and the minimizer $\bar{u}$ of the continuous problem (P). We remark that this is where the second–order results of Section 6.4 enter most strongly.

Theorem 6.6.3. *The following results hold:*

a) Let $\bar{u} \in L^2(0,T)$ be a strict local minimum of $\hat{J}$. Then there exists a sequence of local minima $\bar{u}_K$ of (P_K) such that $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ as $K \to \infty$.

b) For $K \in \mathbb{N}$, let $\bar{u}_K$ be a solution of (P_K). Then there exists $\bar{u} \in C^2([0,T])$ solution of (P) with $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ as $K \to \infty$.

c) In both cases a) and b), the sequence $(\bar{u}_K)$ is bounded in H^1, uniformly in K. Therefore, the constants $C^{(1)}$ and $C^{(2)}$ from Lemma 6.6.1 and 6.6.2 are independent of K.

Proof. We use the following well–known results for the L^2–Hilbert space in part a) and b): Let $f_k \rightharpoonup f$ and $g_k \to g$ both in $L^2(0,T)$ as $k \to \infty$. Then it holds that $\langle f_k, g_k\rangle_2 \to \langle f, g\rangle_2$. Furthermore, $\|f_k\|_2 \to \|f\|_2$ if and only if $f_k \to f$ strongly in $L^2(0,T)$.
We start with the proof of a). First, we denote with $\varepsilon > 0$ the radius given by Definition 6.2.2 for the strict local minimum $\bar{u}$ of $\hat{J}$. Thus, for every $0 < \varepsilon' < \varepsilon$ it holds that

$$\hat{J}(\bar{u}) = \min\left\{\hat{J}(u) \mid u \in \overline{B}_{\varepsilon'}(\bar{u}; L^2)\right\},$$

where we denote by $\overline{B}_{\varepsilon'}(\bar{u}; L^2) = \{u \in L^2(0,T) \mid \|u - \bar{u}\|_{L^2} \leq \varepsilon'\}$ the closed ball around $\bar{u}$ of radius ε' with respect to the $L^2(0,T)$–norm. We construct the discrete local minima sought by considering for $\varepsilon' := \varepsilon/2$ the problem

$$\min\left\{\hat{J}_K(u) \mid u \in \mathcal{P}^2_K \cap \overline{B}_{\varepsilon'}(\bar{u}; L^2)\right\}. \tag{6.37}$$

There exists $N_{\varepsilon'} \in \mathbb{N}$ sufficiently large such that for all $N \geq N_{\varepsilon'}$ we have the estimate $\|\bar{u} - \mathrm{Proj}^2_K(\bar{u})\|_{L^2} < \varepsilon'$. Thus, $\mathcal{P}^2_K \cap \overline{B}_{\varepsilon'}(\bar{u}; L^2)$ is non–empty, and we conclude the existence of a solution $\bar{u}_K$ to (6.37) for every $N \geq N_{\varepsilon'}$.
Notice that at this point, we cannot claim that the solution $\bar{u}_K$ is a local minimum of $\hat{J}_K$ in the sense of Definition 6.2.2, since possibly $\bar{u}_K$ is on the boundary of the ε'–ball from (6.37). However, because $\bar{u}$ is a strict local minimum, in the following we show that $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$, which implies that $\bar{u}_K$ is in the open ball for sufficiently large K and therefore is a local minimum of $\hat{J}_K$. For this purpose, we see that $(\bar{u}_K)_{N \geq N_{\varepsilon'}}$ is bounded in $L^2(0,T)$ by $\|\bar{u}\|_{L^2} + \varepsilon'$ uniformly in K. Therefore, there exists a weak limit w contained in the L^2–closure of the ball, i.e, $w \in \overline{B}_{\varepsilon'}(\bar{u}; L^2)$. If we can show that $\hat{J}(\bar{u}) = \hat{J}(w)$, then the local uniqueness result, see Theorem 6.3.2 i), implies $\bar{u} = w$. To this end, notice that

$$\hat{J}(w) \leq \liminf_{K\to\infty} \hat{J}_K(\bar{u}_K) \leq \limsup_{K\to\infty} \hat{J}_K\left(\mathrm{Proj}^2_K(\bar{u})\right) = \hat{J}(\bar{u}) \leq \hat{J}(w),$$

where we used Lemma 6.6.1 b) for the first estimate. In conclusion, we have shown that $\bar{u}_K \rightharpoonup \bar{u}$ in $L^2(0,T)$ after selecting a subsequence with indices in $S \subset \mathbb{N}$.

Next, we prove convergence of the norms $\|\bar{u}_K\|_2 \to \|\bar{u}\|_2$, from which the required strong L^2–convergence follows. For this aim, recall that $\bar{y}_K := G_K(\bar{u}_K) \to \bar{y} := G(\bar{u})$ and $\bar{q}_K := Q_K(\bar{u}_K) \to \bar{q} := Q(\bar{u})$ in $L^\infty(0,T)$ due to Lemma 6.6.1 b). On the other hand, notice that $\bar{u}$ being a local minimum of $\hat{J}$ and $\bar{u}_K$ being a minimum of $\hat{J}_K$ in $\mathcal{P}^2_K \cap \overline{B}_{\varepsilon'}(\bar{u}; L^2)$ imply the following

$$0 = \hat{J}'(\bar{u})v = \gamma\langle \bar{u}, v\rangle_2 + \langle \bar{q}^\top B\,\bar{y}, v\rangle_2, \quad v \in L^2(0,T), \tag{6.38}$$

$$0 \le \hat{J}'_K(\bar{u}_K)(u-\bar{u}_K) = \gamma\langle \bar{u}_K, u-\bar{u}_K\rangle_2 + \langle \bar{q}_K^\top B\,\bar{y}_K, u-\bar{u}_K\rangle_2, \quad u \in \mathcal{P}^2_K \cap \overline{B}_{\varepsilon'}(\bar{u}; L^2). \tag{6.39}$$

Testing equation (6.38) with $v = \bar{u}$ and inequality (6.39) with $u = \mathrm{Proj}^2_K(\bar{u})$ gives

$$\|\bar{u}\|_2^2 = -\frac{1}{\gamma}\langle \bar{q}^\top B\,\bar{y}, \bar{u}\rangle_2, \quad \|\bar{u}_K\|_2^2 \le \langle \bar{u}_K, \mathrm{Proj}^2_K(\bar{u})\rangle_2 + \frac{1}{\gamma}\langle \bar{q}_K^\top B\,\bar{y}_K, \mathrm{Proj}^2_K(\bar{u}) - \bar{u}_K\rangle_2.$$

This result yields the desired convergence of the norms for $K \in S, K \to \infty$

$$\begin{aligned}\|\bar{u}_K\|_2^2 - \|\bar{u}\|_2^2 \le &\langle \bar{u}_K, \mathrm{Proj}^2_K(\bar{u})\rangle_2 + \frac{1}{\gamma}\langle \bar{q}_K^\top B\,\bar{y}_K, \mathrm{Proj}^2_K(\bar{u})\rangle_2 \\ &- \frac{1}{\gamma}\langle \bar{q}_K^\top B\,\bar{y}_K, \bar{u}_K\rangle_2 + \langle \bar{q}^\top B\,\bar{y}, \bar{u}\rangle_2 \to 0.\end{aligned}$$

The latter scalar products converge since both $q_K^\top B\,\bar{y}_K \to \bar{q}^\top B\,\bar{y}$ and $\mathrm{Proj}^2_K(\bar{u}) \to \bar{u}$ strongly in $L^2(0,T)$ as $K \in S, K \to \infty$. Hence, we have proved $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$.

A consequence of this convergence result is that all $\bar{u}_K$ have to be in the interior of the ball $\mathcal{P}^2_K \cap \overline{B}_{\varepsilon'}(\bar{u}; L^2)$ for all $K \in S$ sufficiently large. Therefore, $\bar{u}_K$ has to be a local minimum of $\hat{J}_K$, i.e., there exists $\varepsilon^* > 0$ such that

$$\hat{J}_K(\bar{u}_K) = \min\left\{\hat{J}_K(u) \mid u \in \mathcal{P}^2_K \cap B_{\varepsilon^*}(\bar{u}_K; L^2)\right\}. \tag{6.40}$$

Next, we prove assertion b) and consider the sequence $(\bar{u}_K)_{K\in\mathbb{N}}$ of solutions of (P_K). First, we need a uniform bound in $L^2(0,T)$ of this sequence. This follows from $\hat{J}_K(\bar{u}_K) \le \hat{J}_K(0)$, after inserting $G(0)$, which simply is the solution of $y' = Ay$ with $y(0) = y_0$. This yields

$$\begin{aligned}\gamma\|u_K\|_2^2 \le &-\beta\|G_K(u_K) - y_{d,K}\|_2^2 - \alpha|G_K(u_K)(T) - y_T|^2 \\ &+ \beta\|G_K(0) - y_{d,K}\|_2^2 + \alpha|G_K(0)(T) - y_T|^2 \\ \le &\beta\|G_K(0) - y_{d,K}\|_2^2 + \alpha|G_K(0)(T) - y_T|^2 \le C_J.\end{aligned}$$

In conclusion, $\|u_K\|_2^2$ is bounded by a constant depending only on given quantities. Hence, there exists a weak limit $w \in L^2(0,T)$. Now, let $\bar{u} \in L^2(0,T)$ be the solution of (P). Analogously to a) we obtain

$$\hat{J}(w) \le \liminf_{K\to\infty} \hat{J}_K(\bar{u}_K) \le \limsup_{K\to\infty} \hat{J}_K\left(\mathrm{Proj}^2_K(\bar{u})\right) = \hat{J}(\bar{u}) \le \hat{J}(w).$$

Notice that, in this argument, it is essential that $\bar{u}_K$ and $\bar{u}$ are not only local but global minima of (P_K) and (P), respectively. This implies that $\bar{u}_K \rightharpoonup \bar{u}$ in $L^2(0,T)$ for a subsequence, which consequently yields $\bar{y}_K \to \bar{y}$ and $\bar{q}_K \to \bar{q}$ in $L^\infty(0,T)$. Due to the fact that $\hat{J}'(\bar{u}) = 0$ on $L^2(0,T)$ and $\hat{J}'_K(\bar{u}_K) = 0$ on $\mathcal{P}^2_K$, we conclude that

$$\begin{aligned}\gamma\|\bar{u}\|_2^2 - \gamma\|\bar{u}_K\|_2^2 &= -\langle \bar{q}^\top B\,\bar{y}, \bar{u}\rangle_2 + \langle \bar{q}_K^\top B\,\bar{y}_K, \bar{u}_K\rangle_2 \\ &= \langle \bar{q}^\top B\,\bar{y}, \bar{u}_K - \bar{u}\rangle_2 + \langle \bar{q}_K^\top B\,\bar{y}_K - \bar{q} B\,\bar{y}, \bar{u}_K\rangle_2 \to 0.\end{aligned}$$

This implies the strong convergence $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ (for a subsequence). Furthermore, both in a) and b), it follows by the standard argument from Lemma 1.4.3 that $\bar{u}_K \to \bar{u}$ holds without selection of a

subsequence. Lastly, recall that local minima of $\hat{J}$ have the higher regularity C^2. This result completes the proof of assertion a) and b).
Part c) follows from the implicit formula for $\bar{u}_K$ stated in Theorem 6.5.2 with analogous estimates as in (6.16). The functions $\bar{y}_K, \bar{q}_K$ are uniformly bounded in the H^1–norm due to the fact that $\bar{u}_K$ is uniformly bounded in the L^2–norm. □

Notice that the parts a) and b) of Theorem 6.6.3 can be proven analogously for (suitable) other discrete spaces for the control. This is due to the fact that we do not exploit the discrete implicit formula for $\bar{u}_K$ and we make no statement about convergence rates.
Based on the L^2–convergence results stated in Theorem 6.6.3, we can prove the following.

Corollary 6.6.4. *Let $\bar{u}$ fulfill the sufficient second–order condition and let $\Lambda, \varepsilon > 0$ be given by Theorem 6.3.2. Furthermore, let $(\bar{u}_K)_{K\in\mathbb{N}} \subset L^2(0,T)$ and $K_\varepsilon \in \mathbb{N}$ such that $\|\bar{u}_K - \bar{u}\|_2 < \varepsilon/2$ for all $K \geq K_\varepsilon$. Then for all $K \geq K_\varepsilon$, it holds that*

$$\frac{\Lambda}{2}\|\bar{u} - \bar{u}_K\|_2^2 \leq \Big(\hat{J}'(\bar{u}_K) - \hat{J}'(\bar{u})\Big)(\bar{u}_K - \bar{u}).$$

Proof. Taking $v = \bar{u}_K - \bar{u}$ in Theorem 6.3.2 iii), it holds that

$$\frac{\Lambda}{2}\|\bar{u} - \bar{u}_K\|_2^2 \leq \hat{J}''(w)\big((\bar{u}_K - \bar{u}), (\bar{u}_K - \bar{u})\big)$$

for every $w \in B_\varepsilon(\bar{u}; L^2)$. Furthermore, by the mean value theorem, for every $K \in \mathbb{N}$ there exists some $\lambda_K \in [0,1]$ such that for $w_K := \lambda_K \bar{u} + (1-\lambda_K)\bar{u}_K$, we have

$$\hat{J}''(w_K)\big((\bar{u}_K - \bar{u}), (\bar{u}_K - \bar{u})\big) = (\hat{J}'(\bar{u}_K) - \hat{J}'(\bar{u}))(\bar{u}_K - \bar{u}).$$

Since $w_K \in B_\varepsilon(\bar{u}; L^2)$ for $K \geq K_\varepsilon$, the claim follows. □

6.7 Accuracy estimates

In this section, we use the results presented above to prove second–order accuracy of the optimal control computed with the finite element method. The following theorem improves the statements of Theorem 6.6.3 and is the second main result of this chapter.

Theorem 6.7.1. *Consider the following two cases:*

a) Let $\bar{u}$ fulfill the sufficient first– and second–order conditions of the minimization problem (P) from Section 6.2, that is $J'(\bar{u})v = 0$ and $\hat{J}''(\bar{u})(v,v) > 0$ for all $v \in L^2(0,T)\backslash\{0\}$. Let $(\bar{u}_K)_{K\in\mathbb{N}}$ be the sequence of local minima of $\hat{J}_K$ given by Theorem 6.6.3 a).

b) For $K \in \mathbb{N}$, let $\bar{u}_K$ be a solution of (P_K) from section (6.5) and let $\bar{u}$ be given by Theorem 6.6.3 b). Furthermore, let $\bar{u}$ fulfil the sufficient second–order condition.

In both cases, there exists $C > 0$ depending (continuously) only on $|A|, |B|, |y_0|, T$ related to the state equation and on $\gamma, \beta, \alpha, \|y_d\|_{H^2}, |y_T|$ related to the cost functional J, such that

$$\|\bar{u} - \bar{u}_K\|_2 \leq CK^{-2}. \tag{6.41}$$

We remark that the rate K^{-2} has to be replaced by K^{-1} if $\sigma = 1$ in (6.29).

Proof. The starting point of the proof is the estimate from Corollary 6.6.4. Notice that $\hat{J}'_K(\bar{u}_K)v = 0$ for all $v \in L^2(0,T)$ by Theorem 6.5.2. Thus, we obtain the estimate

$$\frac{\Lambda}{2}\|\bar{u} - \bar{u}_K\|_2^2 \leq \Big(\hat{J}'(\bar{u}) - \hat{J}'(\bar{u}_K)\Big)(\bar{u} - \bar{u}_K) = \Big(\hat{J}'_K(\bar{u}_K) - \hat{J}'(\bar{u}_K)\Big)(\bar{u} - \bar{u}_K). \tag{6.42}$$

We recall the uniform bound $\|\bar{u}_K\|_{H^1} \leq C_J$ from Theorem 6.6.3 c). Now, we apply Lemma 6.6.2 with $v = \bar{u} - \bar{u}_K$ and conclude that

$$\frac{\Lambda}{2}\|\bar{u} - \bar{u}_K\|_2^2 \leq C_J K^{-2}\|\bar{u} - \bar{u}_K\|_2.$$

This yields the desired second–order accuracy estimate $\|\bar{u} - \bar{u}_K\|_2 \leq C_J\, K^{-2}$. □

6.8 Accuracy estimates with box constraints on the control

In this section, we extend our analysis of accuracy to the case of a finite element approximation to a control–constrained optimal control problem. For the case where the controls are assumed piecewise constant, we prove the estimate $\|\bar{u} - \bar{u}_K\|_2 \leq CK^{-\sigma}$ with $\sigma = 1$, whereas in the case of continuous piecewise quadratic controls we obtain $\sigma = 2$ if we assume second–order accuracy in (6.29).

We focus on the optimal control problem discussed in the previous sections with the addition of bilateral box constraints that define the following set of admissible controls

$$U_{\text{ad}} := \{u \in L^2(0,T) \mid a \leq u(t) \leq b \quad \text{f.a.e.} \;\; t \in]0,T[\,\}$$

with $-\infty \leq a < b \leq \infty$.

Therefore, based on our construction of the control–to–state map and of the cost functional, our optimal control problem is defined as follows

$$\min_{u \in U_{\text{ad}}} \hat{J}(u), \tag{P_{ad}}$$

The first–order necessary condition for a local minimum $\bar{u} \in U_{\text{ad}}$ becomes a variational inequality of the form

$$J'(\bar{u})(v - \bar{u}) \geq 0, \quad v \in U_{\text{ad}}. \tag{6.43}$$

Consequently, Corollary 6.2.5 must be adapted. The reason is that $\bar{u}$ is in general only Lipschitz continuous and satisfies the following equation

$$\bar{u}(t) = \max\left\{a, \min\left\{b, -\frac{1}{\gamma}\bar{q}(t)^\top B\, \bar{y}(t)\right\}\right\}, \quad t \in [0,T]. \tag{6.44}$$

Clearly, some steps in our analysis cannot be directly performed in this case. Specifically, one cannot transfer the proof to deduce (6.44) from (6.43) for the discrete case with piecewise quadratic controls, and an implicit equation for the discrete local minima like in (6.33) from Theorem 6.5.2 seems difficult to obtain.

One can also see that the inequality condition (6.43) impacts our accuracy analysis, in the sense that the local coercivity $\hat{J}''(u)(v,v) \geq \frac{\Lambda}{2}\|v\|_2^2$ does no longer hold for all $v \in L^2(0,T)$ but on a subset $C_{\bar{u}}^{\tau}$ given below. Consequently, since we set $v = \bar{u}_K - \bar{u}$ in the proof of Corollary 6.6.4, we have to verify that the sequence $(\bar{u}_K - \bar{u})$ is in $C_{\bar{u}}^{\tau}$ for K sufficiently large.

With these preparatory comments, we start our discussion of the optimal control problem (P_{ad}). We remark that all properties of the control–to–state and control–to–adjoint map remain valid, as well as

Theorem 6.2.1 which states the existence of an optimal control. The definition of a (local) minimum of (P_{ad}) is the same as in Definition 6.2.2 with the addition that $\bar u \in U_{\mathrm{ad}}$ and the set $B_\varepsilon(\bar u; L^2)$ has to be replaced by $U_{\mathrm{ad}} \cap B_\varepsilon(\bar u; L^2)$. Next, equation (6.14) of the optimality system has to be adapted. For this purpose, let us focus on a local minimum $\bar u$ of (P_{ad}) and define

$$\Phi(t) := \gamma \bar u(t) + Q(\bar u)(t)^\top B\, G(\bar u)(t), \quad t \in [0,T]. \tag{6.45}$$

By testing (6.43) with proper $v \in U_{\mathrm{ad}}$ one obtains for $t \in [0,T]$ the following implications that give rise to the formula for $\bar u$ given in (6.44):

$$\begin{cases} \Phi(t) > 0 & \Longrightarrow\ \bar u(t) = a, \\ \Phi(t) < 0 & \Longrightarrow\ \bar u(t) = b, \\ a < \bar u(t) < b & \Longrightarrow\ \Phi(t) = 0. \end{cases} \tag{6.46}$$

Next, concerning the accuracy analysis, we have that Lemma 6.3.1 remains valid if $v_k, v \in U_{\mathrm{ad}}$. In order to reformulate Theorem 6.3.2, we follow [21] and define the following sets for a local minima $\bar u \in U_{\mathrm{ad}}$ and $\tau \geq 0$. We have

$$\begin{aligned} S_{\bar u} &:= \{\lambda(v - \bar u) \mid \lambda > 0, v \in U_{\mathrm{ad}}\}, \\ C_{\bar u}^\tau &:= \overline{S_{\bar u}}^{L^2} \cap \left\{ v \in L^2(0,T) \mid |\hat J'(\bar u)v| \leq \tau \|v\|_2 \right\}, \\ E_{\bar u}^\tau &:= \left\{ v \in L^2(0,T) \mid v(t) \begin{cases} \geq 0, & \text{if } \bar u(t) = a \\ \leq 0, & \text{if } \bar u(t) = b \\ = 0, & \text{if } |\Phi(t)| > \tau \end{cases} \right\}. \end{aligned} \tag{6.47}$$

For our bilinear optimal control problem, it can be shown with standard techniques that $E_{\bar u}^0 = C_{\bar u}^0$ since $\gamma > 0$. The sufficient second–order condition for a local minimum $\bar u \in U_{\mathrm{ad}}$ now reads

$$\hat J''(\bar u)(v,v) > 0, \quad v \in C_{\bar u}^0 \backslash \{0\}. \tag{6.48}$$

This fact implies the local coercivity of $\hat J$, in the sense that there exists $\delta, \varepsilon, \Lambda, \tau > 0$ such that

$$\hat J''(u)(v,v) \geq \frac{\Lambda}{2} \|v\|_2^2, \quad v \in C_{\bar u}^\tau,\ u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar u; L^2). \tag{6.49}$$

For later use, we remark that $E_{\bar u}^\tau \subset C_{\bar u}^{\tau\sqrt{T}}$, since for $v \in E_{\bar u}^\tau$ it holds that

$$|\hat J'(\bar u)v| = \int_{\{t \in [0,T] \mid |\Phi(t)| < \tau\}} |\Phi(t) v(t)|\, dt \leq \tau \int_0^T |v(t)|\, dt \leq \tau \sqrt{T} \|v\|_2. \tag{6.50}$$

This remark concludes our discussion of the continuous problem with box constraints.

Next, we investigate the discrete finite element approximation to our control–constrained optimal control problem. We see that the properties of the discrete control–to–state map and control–to–adjoint map (6.28)–(6.30), as well as the definition of the reduced cost functional (6.31) remain mostly unchanged. Analogously to the continuous case, Theorem 6.5.2 and the implicit equation for a local minimum $\bar u_K$ have to be adapted as discussed below.

We start with the case of piecewise constant controls in the space

$$\mathcal{P}_K^0 = \{\psi : [0,T[\to \mathbb{R} \mid \psi \text{ is constant on } [t_i, t_{i+1}[\,, i = 0, \ldots, K-1\} \subset L^2(0,T).$$

For a local minimum $\bar u_K \in U_{\mathrm{ad}} \cap \mathcal{P}_K^0$, we define

$$\Phi_K(t) := \gamma \bar u_K(t) + Q_K(\bar u_K)(t)^\top B\, G_K(\bar u_K)(t), \quad t \in [0,T], \tag{6.51}$$

and the Fréchet differentiabilty of $\hat{J}_K$ implies the necessary condition $\hat{J}_K'(\bar{u}_K)(v-\bar{u}_K) \geq 0$ for all $v \in U_{\mathrm{ad}} \cap \mathcal{P}_K^0$. The choice of piecewise constant controls allows us to test with the indicator function on the interval $[t_i, t_{i+1}[$

$$t \mapsto v(t) = 1_{|[t_i,t_{i+1}[}(t) \in \mathcal{P}_K^0.$$

Thus, it holds for all $i = 0, \dots, K-1$ and all numbers $x \in [a, b]$ that

$$\int_{t_i}^{t_{i+1}} \Phi_K(t)\, dt \left(x - \bar{u}_K^i\right) \geq 0,$$

where we use the notation $\bar{u}_K^i := \bar{u}_K(t_i)$.

Next, we introduce the following mean values within one discrete time interval

$$s_i := \frac{1}{\Delta t} \int_{t_i}^{t_{i+1}} -\frac{1}{\gamma} Q_K(\bar{u}_K)(t)^\top B\, G_K(\bar{u}_K)(t)\, dt, \qquad i = 0, \dots, K-1.$$

With this construction, we obtain

$$\begin{cases} \bar{u}_K(t_i) = a & \Longrightarrow \int_{t_i}^{t_{i+1}} \Phi_K(t)\, dt \geq 0 \Longrightarrow \bar{u}_K^i \geq s_i, \\ \bar{u}_K(t_i) = b & \Longrightarrow \int_{t_i}^{t_{i+1}} \Phi_K(t)\, dt \leq 0 \Longrightarrow \bar{u}_K^i \leq s_i, \\ a < \bar{u}_K(t_i) < b & \Longrightarrow \int_{t_i}^{t_{i+1}} \Phi_K(t)\, dt = 0 \Longrightarrow \bar{u}_K^i = s_i. \end{cases} \tag{6.52}$$

This result yields the implicit formula $\bar{u}_K(t_i) = \max\{a, \min\{b, s_i\}\}$; compare this with (6.44). Next, concerning the accuracy of G_K and Q_K in the case of piecewise constant controls with box constraints, in view of (6.25), we assume first–order estimates, that is for every $u \in \mathcal{P}_K^0$ there exists $C = C_{\|u\|_\infty} > 0$ such that

$$\begin{aligned} &\|G_K(u) - G(u)\|_2 + \|Q_K(u) - Q(u)\|_2 \leq CK^{-1}, \qquad K \in \mathbb{N}, \qquad (6.53)\\ &G_K(u) \to G(u), \quad Q_K(u) \to Q(u) \qquad \text{in } H^1 \text{ and uniformly on } [0,T] \text{ as } K \to \infty. \end{aligned}$$

Now, concerning our analysis in Section 6.6, we see that Lemma 6.6.1 remains valid if we replace the rates K^{-2} with K^{-1}.

Further, due to the implicit formula (6.52), it can be shown that $\bar{u}_K \to \bar{u}$ uniformly on $[0, T]$. Theorem 6.6.3 also remains true; its proof changes only slightly.

The next issue is to verify Corollary 6.6.4. To this end, we need to make sure that $v = \bar{u}_K - \bar{u} \in C_{\bar{u}}^\tau$ in order to apply the local coercivity of $\hat{J}''$. Due to $E_{\bar{u}}^{\tau'} \subset C_{\bar{u}}^\tau$ with $\tau' := \tau/\sqrt{T}$, it is sufficient to prove that for all $t \in [0, T]$ it holds

$$\bar{u}_K(t) \begin{cases} \geq \bar{u}(t), & \text{if } \bar{u}(t) = a, \\ \leq \bar{u}(t), & \text{if } \bar{u}(t) = b, \\ = \bar{u}(t), & \text{if } |\Phi(t)| > \tau'. \end{cases}$$

The first two claims are clearly fulfilled since $\bar{u}_K \in U_{\mathrm{ad}}$. The third claim is fulfilled for K sufficiently large and can be shown as follows. Notice that $\Phi_K \to \Phi$ uniformly on $[0, T]$ due to the properties of G_K and Q_K and since $\bar{u}_K \to \bar{u}$ uniformly. Further, define the set $M_{\tau'} := \{t \in [0, T] \mid |\Phi(t)| > \tau'\}$ and notice with (6.46) that $\bar{u}(t) \in \{a, b\}$ for $t \in M_{\tau'}$. Thus, we fix any $t' \in M_{\tau'}$ with $t' \in [t_j, t_{j+1}[$ and want to prove $\int_{t_j}^{t_{j+1}} \Phi_K(t)\, dt \neq 0$ for K sufficiently large as this implies $\bar{u}_K(t') \in \{a, b\}$ due to (6.52). Due to the Lipschitz continuity of Φ and the above mentioned uniform convergence, there exists $K_0 \in \mathbb{N}$ such that for all $K \geq N_0$ and all $t \in [t_j, t_{j+1}[$ it holds

$$\Phi(t') - \Phi(t) \leq L\Delta t = LT/K \leq \tau'/4 \quad \text{and} \quad \Phi_K(t) - \Phi(t) \leq \tau'/4.$$

Because $t' \in M_{\tau'}$ it holds that $\Phi(t') > \tau$ or $\Phi(t') < -\tau$. In the first case, we find that

$$\int_{t_j}^{t_{j+1}} \Phi_K(t)\,dt \geq \int_{t_j}^{t_{j+1}} \Big(-|\Phi_K(t) - \Phi(t)| - |\Phi(t) - \Phi(t')| + \Phi(t').\Big)\,dt \geq -\Delta t\,\tau'/2 + \Delta t\,\Phi(t') > 0.$$

We argue analogously for the second case and conclude that $\int_{t_j}^{t_{j+1}} \Phi_K(t)\,dt \neq 0$. This completes the proof of $\bar{u}_K - \bar{u} \in C_{\bar{u}}^{\tau}$.

Now, we continue our discussion referring to Section 6.7. In the presence of box constraints, it is important to construct some discrete control $w_K \in \mathcal{P}_K^0$ such that

$$\hat{J}'(\bar{u})\bar{u} = \hat{J}_K'(\bar{u}_K)w_K.$$

It can be shown that the choice

$$w_K(t) := \begin{cases} \left(\int_{t_j}^{t_{j+1}} \Phi(s)\,ds\right)^{-1} \int_{t_j}^{t_{j+1}} \bar{u}(s)\Phi(s)\,ds, & \int_{t_j}^{t_{j+1}} \Phi(s)\,ds \neq 0 \\ \Delta t^{-1} \int_{t_j}^{t_{j+1}} \bar{u}(s)\,ds, & \int_{t_j}^{t_{j+1}} \Phi(s)\,ds = 0, \end{cases} \tag{6.54}$$

for $j = 0, \dots, K-1$, $t \in [t_j, t_{j+1}[$ fulfills $\hat{J}'(\bar{u})(\bar{u} - w_K) = 0$ with linear accuracy $\|\bar{u} - w_K\|_2 \leq CK^{-1}$. Thus, using this result, we obtain the linear error estimate $\|\bar{u} - \bar{u}_K\|_2 \leq C_J K^{-1}$ with Corollary 6.6.4 and Young's inequality as follows

$$\begin{aligned} \frac{\Lambda}{2}\|\bar{u} - \bar{u}_K\|_2^2 &\leq \Big(\hat{J}'(\bar{u}) - \hat{J}'(\bar{u}_K)\Big)(\bar{u} - \bar{u}_K) \\ &\leq \Big(\hat{J}_K'(\bar{u}_K) - \hat{J}'(\bar{u}_K)\Big)(\bar{u} - \bar{u}_K) + \Big(\hat{J}'(\bar{u}) - \hat{J}_K'(\bar{u}_K)\Big)(\bar{u} - w_K) \\ &\leq CK^{-1}\|\bar{u} - \bar{u}_K\|_2 + C\left(K^{-1} + \|\bar{u} - \bar{u}_K\|_2\right)\|w_K - \bar{u}\|_2 \\ &\leq C\left(K^{-2} + K^{-1}\|w_K - \bar{u}\|_2 + C_\epsilon\|w_K - \bar{u}\|_2^2 + \epsilon\|\bar{u} - \bar{u}_K\|_2^2\right). \end{aligned} \tag{6.55}$$

Next, we present our analysis of the case of continuous, piecewise quadratic discrete controls. Analogously to Lemma 6.5.1, we introduce for $K \in \mathbb{N}$ the semidiscrete optimal control problem, that is, we discretize the control–to–state map G and the target state in J but not the set of admissible controls:

$$\min_{u \in U_{\mathrm{ad}}} \hat{J}_K(u) \quad \text{with } \hat{J}_K \text{ defined as in (6.31).} \tag{P_{ad}^K}$$

The necessary condition for local minima of $\hat{J}_K$ now yields the analogous result to Theorem 6.5.2.

Theorem 6.8.1. *Let $\bar{u}_K \in U_{\mathrm{ad}}$ be a local solution to (P_{ad}^K) and define $\bar{y}_K := G_K(\bar{u}_K)$, $\bar{q}_K := Q_K(\bar{u}_K) \in (\mathcal{P}_K^1)^N$. Then the implicit formula*

$$\bar{u}_K(t) = \max\left\{a, \min\left\{b, -\frac{1}{\gamma}\bar{q}_K(t)^\top B\,\bar{y}_K(t)\right\}\right\}, \quad t \in [0, T] \tag{6.56}$$

holds and $\bar{u}_K$ is a continuous piecewise quadratic polynomial (not necessarily on the same, uniform grid).

Analogously to (P_K), the problem (P_{ad}^K) is a-priori not a finite–dimensional one, however, equation (6.56) implies that calculating local minima is a finite–dimensional problem. Thus in application, it is numerically possible to compute the grid values of $\bar{u}_K$ and one can follow the approach from Section 9. The next Lemma is the analogue to Theorem 6.6.3.

Lemma 6.8.2. *The following results hold:*

a) *Let $\bar{u} \in U_{\mathrm{ad}}$ be a strict local minimum of $\hat{J}$ over U_{ad}. Then there exists a sequence of local minima $\bar{u}_K$ of (P_{ad}^K) such that $\bar{u}_K \to \bar{u}$ in $L^2(0, T)$ as $K \to \infty$.*

b) For $K \in \mathbb{N}$ let $\bar{u}_K$ be a solution of (P_{ad}^K). Then there exists a Lipschitz continuous function $\bar{u}$ solution of (P_{ad}) with $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ as $K \to \infty$.

c) In both cases a) and b), the sequence $(\bar{u}_K)$ is bounded in H^1, uniformly in K.

Proof. We only sketch the proof due to its similarity to the proof of Theorem 6.6.3. In a), the sequence $(\bar{u}_K)$ is constructed analogously. Thus, in both cases a) and b), the weak convergence can be shown in the same manner. Lastly, the strong L^2–convergence of $\bar{u}_K$ to $\bar{u}$ follows from (6.56) and Lemma 6.6.1 b). For part c), recall that both $\max\{f,g\}, \min\{f,g\} \in W^{1,p}(0,T)$ if $f, g, \in W^{1,p}(0,T)$ and their $W^{1,p}$–norm is bounded by $\|f\|_{W^{1,p}} + \|g\|_{W^{1,p}}$. □

The next step is to prove Corollary 6.6.4. It is only left to show that $v = \bar{u}_K - \bar{u}$ is in $C_{\bar{u}}^{\tau}$ for sufficiently large K. Once again, this is a consequence of the representation (6.56) and we omit the details.
The difference to (6.55) is that $\hat{J}_K'(\bar{u}_K)(v - \bar{u}_K) \geq 0$ holds for all $v \in U_{\text{ad}}$ instead of v from a discrete space. Therefore, we may simply set $w_K = \bar{u}$, and an analogous coercivity estimate from Corollary 6.6.4 yields

$$\begin{aligned}\frac{\Lambda}{2}\|\bar{u} - \bar{u}_K\|_2^2 &\leq \left(\hat{J}'(\bar{u}) - \hat{J}'(\bar{u}_K)\right)(\bar{u} - \bar{u}_K) \leq \left(\hat{J}_K'(\bar{u}_K) - \hat{J}'(\bar{u}_K)\right)(\bar{u} - \bar{u}_K) \\ &\leq C_J K^{-1}\|\bar{u} - \bar{u}_K\|_2.\end{aligned} \tag{6.57}$$

The last estimate follows from an application of Lemma 6.6.2.
This concludes the discussion of the box-constrained optimal control problem, and we summarize the results in the following two theorems; compare this with Theorem 6.6.3 and Theorem 6.7.1.

Theorem 6.8.3. *(The piecewise constant case with box constraints)*
The following results hold.

a) Let $\bar{u} \in U_{\text{ad}}$ be a local minimum of $\hat{J}$, which fulfills the second–order condition from Theorem 6.3.2. Then there exists a sequence of local minima $\bar{u}_K \subset U_{\text{ad}} \cap \mathcal{P}_K^0$ of the corresponding discrete optimal control problem such that $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ as $K \to \infty$.

b) For $K \in \mathbb{N}$, let $\bar{u}_K \in U_{\text{ad}} \cap \mathcal{P}_K^0$ be a solution of the discrete optimal control problem. Then there exists $\bar{u} \in U_{\text{ad}}$ solution of (P) with $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ as $K \to \infty$.

In both cases, we have the linear accuracy

$$\|\bar{u}_K - \bar{u}\|_2 \leq CK^{-1}.$$

Theorem 6.8.4. *(The continuous piecewise quadratic case with box constraints)*
The following results hold.

a) Let $\bar{u} \in U_{\text{ad}}$ be a local minimum of $\hat{J}$ which fulfills the second–order necessary condition. Then there exists a sequence of local minima $\bar{u}_K \subset U_{\text{ad}}$ of the corresponding discrete optimal control problem

$$\min_{u \in U_{\text{ad}}} \hat{J}_K(u), \tag{6.58}$$

where $\bar{u}_K$ is a continuous, piecewise polynomial and $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ as $K \to \infty$.

b) For $K \in \mathbb{N}$, let $\bar{u}_K \in U_{\text{ad}}$ be a solution of the discrete optimal control problem (6.58). Then there exists $\bar{u} \in U_{\text{ad}}$ solution of (P) with $\bar{u}_K \to \bar{u}$ in $L^2(0,T)$ as $K \to \infty$.

In both cases, we have at least first–order accuracy

$$\|\bar{u}_K - \bar{u}\|_2 \leq CK^{-1}.$$

Assuming $\sigma = 2$ in (6.29), we remark that second–order accuracy holds in Theorem 6.8.4.

6.9 Numerical approximation and optimization

In this section, numerical evidence is presented that supports our theoretical findings in the unconstrained case. More precisely, we see quadratic accuracy of the finite element approximation (7.6)–(7.7) for the forward and backward problem. Furthermore, Theorem 6.7.1 is verified numerically in the sense that finite–dimensional solutions $\bar{u}_K$ of (P_K) are computed and the convergence (6.41) with quadratic accuracy to an exact solution $\bar{u}$ of (P) holds.

We start this section, by calculating the finite element scheme for $i = 0, \dots, K$

$$\langle y_K', \psi_i \rangle_2 = \langle (A + uB) y_K, \psi_i \rangle_2, \tag{6.59}$$

$$-\langle q_K', \psi_i \rangle_2 = \langle (A + uB)^\top q_K + \beta(y_K - y_{d,K}), \psi_i \rangle_2, \tag{6.60}$$

using the hat functions $\{\psi_i \mid i = 0, \dots, K\}$ given in (6.22)–(6.23) as basis of $\mathcal{P}_K^1$. Let $K \in \mathbb{N}$ be arbitrary but fixed. In the following computations, we write an upper index i for an evaluation at time point $t_i = i\Delta t$ and we omit to write the lower index for the accuracy. Thus, $y^i := y_K(t_i)$, $q^i := q_K(t_i)^\top$ and $u^i = u_K(t_i)$. Notice that the components of y_K and q_K belong to $\mathcal{P}_K^1$ and hence on an interval $[t_i, t_{i+1}]$ we can use the (linear Taylor) representation

$$y_K(t) = y^i + (t - t_i) \frac{y^{i+1} - y^i}{\Delta t}, \quad t \in [t_i, t_{i+1}]; \tag{6.61}$$

an analogous representation holds for q_K. In order for $\bar{u}_K \in \mathcal{P}_K^2$ to be determined on a subintervall $[t_i, t_{i+1}[$, we introduce the midpoints

$$t_{i+1/2} := t_i + \Delta t/2, \quad t_{i-1/2} := t_i - \Delta t/2 \qquad \text{for } i = 0, \dots, K.$$

Thus, the exact (quadratic Taylor) representation of a discrete control $u_K \in \mathcal{P}_K^2$ is given by

$$u_K(t) = u^i + \frac{t - t_i}{\Delta t}\left(-u^{i+1} + 4u^{i+1/2} - 3u^i\right) + \frac{(t - t_i)^2}{\Delta t^2} 2 \left(u^{i+1} - 2u^{i+1/2} + u^i\right), \quad t \in [t_i, t_{i+1}].$$

In the numerical implementation, the piecewise quadratic polynomial $\bar{u}_K$ will be computed via the discrete reduced gradient (6.33). Therefore, the evaluation of u_K at the intermediate points is a numerical available procedure; recall that due to the finite element approximation of y_K and q_K, we obtain the (exact) intermediate values of the discrete state and discrete co–state simply by

$$y_K(t_{i+1/2}) = \frac{y^{i+1} + y^i}{2} \quad \text{and} \quad q_K(t_{i+1/2}) = \frac{q^{i+1} + q^i}{2}.$$

For the forward problem (6.59), the left–hand side becomes

$$\langle y_K', \psi_i \rangle_2 = \frac{1}{2}\left(y^{i+1} - y^{i-1}\right), \qquad i = 1, \dots K - 1,$$

$$\langle y_K', \psi_0 \rangle_2 = \frac{1}{2}\left(y^1 - y^0\right), \quad \langle y_K', \psi_K \rangle_2 = \frac{1}{2}\left(y^K - y^{K-1}\right).$$

The right–hand side reads for $u_K \in \mathcal{P}_K^2$, $i = 1, \dots, K - 1$

$$\begin{aligned}\langle (A + u_K B) y_K, \psi_i \rangle_2 = \frac{\Delta t}{60}\Big(& \big(10A + (u^{i-1} + 8u^{i-1/2} + u^i)B\big) y^{i-1} \\ & + \big(40A + (-u^{i-1} + 12u^{i-1/2} + 18u^i + 12u^{i+1/2} - u^{i+1})B\big) y^i \\ & + \big(10A + (u^i + 8u^{i+1/2} + u^{i+1})B\big) y^{i+1}\Big).\end{aligned}$$

Let $I_N \in \mathbb{R}^{N\times N}$ denote the identity matrix. Combining both results, we obtain the following implicit scheme: The value $y^0 \in \mathbb{R}^N$ is given by the initial condition and the first step to calculate y^1 is

$$\Big(I_N - \frac{\Delta t}{30}\big(10A + (u^0 + 8u^{1/2} + u^1)B\big)\Big)y^1 = \Big(I_N + \frac{\Delta t}{30}\big(20A + (9u^0 + 12u^{1/2} - u^1)B\big)\Big)y^0.$$

For $i = 1, \ldots, K-2$, the value of y^{i+1} is computed from y^i and y^{i-1} via

$$\begin{aligned} y^{i+1} - y^{i-1} = \frac{\Delta t}{30}\Big(&\big(10A + (u^{i-1} + 8u^{i-1/2} + u^i)B\big)y^{i-1} \\ &+\big(40A + (-u^{i-1} + 12u^{i-1/2} + 18u^i + u^{i+1/2} + u^{i+1})B\big)y^i \\ &+\big(10A + (u^i + 8u^{i+1/2} + u^{i+1})B\big)y^{i+1}\Big). \end{aligned} \tag{6.62}$$

Notice that this scheme is well–defined in the sense that we can solve for y^{i+1} if Δt is sufficiently small depending on $A, B, \|u\|_\infty$. Once again, we use the uniform bound $\|\bar{u}_K\|_{H^1} \leq C_J$.
Next, we consider the accuracy of this scheme and verify (6.28). For this purpose, define

$$F(t) := A + u(t)B \quad \text{and write } F^i := F(u(t_i)) \text{ for } i = 0, \ldots, K.$$

Lemma 6.9.1. *For the problem $y' = (A + u(t)B)y$ with $u \in C^2([0,T])$ and unique solution $y \in C^3([0,T])$, consider the implicit two–step scheme*

$$y^{i+1} - y^{i-1} = \frac{\Delta t}{3}\Big(F^{i-1}y^{i-1} + 4F^i y^i + F^{i+1}y^{i+1}\Big). \tag{6.63}$$

Then, this method is of order 2, that is, it holds for the local error

$$|y(t_i) - y^i| \leq C\Delta t^3 \quad \text{as } \Delta t \to 0.$$

Furthermore, it is stable and consistent, and therefore convergent of order 2.

Proof. The claim follows by an application of Theorem 2.4 with coefficients

$$\alpha_2 = 1,\ \alpha_1 = 0,\ \alpha_0 = -1, \quad \beta_2 = \beta_0 = \frac{1}{3},\ \beta_1 = \frac{4}{3}$$

and Theorem 4.5 from [40, Chapter III]. □

We remark that the scheme (6.63) is known as the Milne–method [40, Chapter III.1]. It has the same order as (6.62) since the difference of the coefficients, e.g., $u^{i-1} + 8u^{i-1/2} + u^i = 10u^{i-1} + \mathcal{O}(\Delta t)$ is of lower order.
Next, we prove (6.29) for linear accuracy $\sigma = 1$.

Lemma 6.9.2. *There exists $K_0 \in \mathbb{N}$ and $C > 0$ such that for all $K \geq K_0$*

$$\|G_K(u_K) - G(u_K)\|_2 + \|Q_K(u_K) - Q(u_K)\|_2 \leq C_{\|u_K\|_{H^1}} K^{-1}.$$

Proof. First, notice that K_0 is given in Lemma 6.4.1 in order to have well–defined functions G_K, Q_K. For $i = 0, \ldots, K$, let

$$y_K := G_K(\bar{u}_K), \quad y := G(\bar{u}_K), \quad z := \mathrm{Proj}_K^1(y) \quad \text{and} \quad y^i := y_K(t_i),\ z^i := z(t_i).$$

Due to the second–order accuracy of the projection (6.24), we obtain

$$\|z - y\|_2 \leq C\|G(\bar{u}_K)\|_{H^2}K^{-2} \leq C\|\bar{u}_K\|_{H^1}K^{-2} \leq C_J K^{-2}.$$

Thus, it is sufficient to prove $\|y_K - z\|_2 \leq CK^{-1}$. We do this by following the lines of [40, Chapter III.2] but instead of considering the pointwise local error, one needs to analyze a L^2–local error: Let y^{i+1} be the numerical solution of (6.63) under the assumption that the exact starting values were used, i.e., $y^i = y(t_i)$ and $y^{i-1} = y(t_{i-1})$. Then the local L^2–error is defined as

$$\|y_K - z\|_{2,i} := \left(\int_{t_i}^{t_{i+1}} |y_K(s) - z(s)|^2 \, ds\right)^{1/2}.$$

Notice that z and the exact solution y coincide at the grid points. Hence, under the assumption that y^i, y^{i-1} are exact and by writing y_K and z via (6.61), we obtain the following

$$\begin{aligned}\|y_K - z\|_{2,i}^2 &= \frac{\Delta t}{3}\Big((y^i - z^i)^2 + (y^{i+1} - z^{i+1})^2 + y^{i+1}(y^i - z^i) + z^{i+1}(y^i - z^i)\Big)\\ &= \frac{\Delta t}{3}(y^{i+1} - z^{i+1})^2 = \frac{\Delta t}{3}(y^{i+1} - y(t_{i+1}))^2.\end{aligned}$$

Thus, the L^2–local error and the pointwise local error are (up to the factor $(\Delta t)^{1/2}$) the same and from [40, Chapter III.2], we can repeat the proof of Lemma 2.2, Theorem 2.4 with $|y'(t) - y'(s)| \leq C_{\|y\|_{H^2}} |t - s|^{1/2}$ to obtain that the local L^2–error is of first-order, that is,

$$\|y_K - z\|_{2,i} \leq C\Delta t^2$$

Lastly, we follow the lines of the proof of Theorem 4.5 to derive the linear global error estimate

$$\|y_K - z\|_{L^2(0,T)} \leq C\Delta t.$$

This proves the first–order accuracy for G_K, and the case for Q_K is done analogously. □

We proceed similarly for the adjoint problem. The left–hand side is

$$\begin{aligned}-\langle q_K', \psi_i\rangle_2 &= \frac{1}{2}\left(q^{i-1} - q^{i+1}\right), \qquad i = 1, \dots K-1,\\ -\langle q_K', \psi_0\rangle_2 &= \frac{1}{2}\left(q^0 - q^1\right), \quad -\langle q_K', \psi_K\rangle_2 = \frac{1}{2}\left(q^{K-1} - q^K\right).\end{aligned}$$

For given $u_K \in \mathcal{P}_K^2$, $y_K \in (\mathcal{P}_K^1)^N$, the terms on the right–hand side become

$$\begin{aligned}\langle q_K^\top(A + uB), \psi_i\rangle_2 =& \frac{\Delta t}{30}\Big(q^{i-1}\big(10A + (u^{i-1} + 8u^{i-1/2} + u^i)B\big)\\ &+ q^i\big(40A + (-u^{i-1} + 12u^{i-1/2} + 18u^i + 12u^{i+1/2} - u^{i+1})B\big)\\ &+ q^{i+1}\big(10A + (u^i + 8u^{i+1/2} + u^{i+1})B\big)\Big),\end{aligned}$$

$$\langle (y_K - y_{d,K}), \psi_i\rangle_2 = \frac{\Delta t}{6}\Big((y^{i-1} - y_d^{i-1}) + 4(y^i - y_d^i) + (y^{i+1} - y_d^{i+1})\Big).$$

The value of q^K is given by the terminal condition, and we arrive at the following two–step scheme

$$\begin{aligned}q^{K-1}\Big(I_K - \frac{\Delta t}{30}\big(10A + (u^K + 8u^{K-1/2} + u^{K-1})B\big)\Big)&\\ =q^K\Big(I_K + \frac{\Delta t}{30}\big(20A + (9u^K + 12u^{K-1/2} - u^{K-1})B\big)\Big)&\\ + \frac{\Delta t}{3}\beta\Big(2(y^K - y_d^K) + y^{K-1} - y_d^{K-1}\Big),&\end{aligned}$$

$$\begin{aligned} q^{i-1}\Big(I_n - \frac{\Delta t}{30}\big(10A{+}(u^{i-1}+8u^{i-1/2}+u^{i})B\big)\Big) \\ =&q^{i+1}\Big(I_n + \frac{\Delta t}{30}\big(10A+(-u^{i-1}+12u^{i-1/2}+18u^{i}+12u^{i+1/2}-u^{i+1})B\big)\Big) \\ &+\frac{\Delta t}{30}q^{i}\Big(40A+(-u^{i-1}+12u^{i-1/2}+18u^{i}+12u^{i+1/2}-u^{i+1})B\Big) \\ &+\frac{\Delta t}{3}\beta\Big(y^{i-1}-y_d^{i-1}+4(y^{i}-y_d^{i})+y^{i+1}-y_d^{i+1}\Big). \end{aligned}$$

We remark that the linear accuracy result for the adjoint can be proven analogously to the case of the state equation.

Once y_K and q_K are computed on $[0,T]$, we can calculate the reduced gradient via

$$\hat{J}_K'(u)(t) = q_K(t)^\top B\, y_K(t) + \gamma u(t), \quad t\in[0,T]. \tag{6.64}$$

Notice that (6.64) is a local formula which is very advantageous from a numerical point of view in the sense that only the values of q_K, y_K and u at time t_i are needed to calculate the gradient at time t_i. We remark that in the case of using other finite element space than $\mathcal{P}_K^2$ for the discrete controls, this feature is lost. For example, taking $\mathcal{P}_K^1$ as the discrete space for the controls, one obtains that the discrete reduced gradient at t_i depends on all grid values of $y_K(t_j), q_K(t_j)$, $j=0,\dots,N$.

Next, we construct an optimal control problem for $N=1$ to which an exact solution can be obtained. First, one picks any (by hand integrable) $\bar{u}\in C^2([0,1])$ and $y_0, A, B\in\mathbb{R}$; we choose $y_0=A=B=1$ and

$$\bar{u}(t) := 2\pi\sin(2\pi t).$$

Correspondingly, we obtain the solution to the forward problem

$$\bar{y}(t) = \exp\left(At + B\int_0^t \bar{u}(\tau)\,d\tau\right) = e^t e^{1-\cos(2\pi t)}.$$

Now, the optimality condition $0=\gamma\bar{u}(t)+\bar{q}(t)^\top B\,\bar{y}(t)$ yields the backward solution

$$\bar{q}(t) = -\frac{\gamma}{B}\frac{\bar{u}(t)}{\bar{y}(t)}, \quad \bar{q}(T)=0.$$

Once $(\bar{u},\bar{y},\bar{q})$ is determined, we compute the components of the cost functional from the backward problem. We choose $\alpha=0, \beta=1$ and obtain the desired state

$$y_d(t) = \bar{q}'(t)+\bar{q}(t)\big(A+\bar{u}(t)B\big)+\bar{y}(t) = -\frac{\gamma}{B}\frac{\bar{u}'(t)}{\bar{y}(t)}+\bar{y}(t).$$

This concludes the construction of our test problem.

Next, we briefly describe the numerical algorithm that solves our optimal control problem. We set the number of grid points K, define a uniform grid on $[0,T]$ and set an initial guess $u_{K,0}:=0\in\mathcal{P}_K^2$ for the discrete optimal control $\bar{u}_K$. The finite–dimensional minimization problem is solved with a non–linear conjugate gradient (NCG) scheme, see Algorithm 1 below.

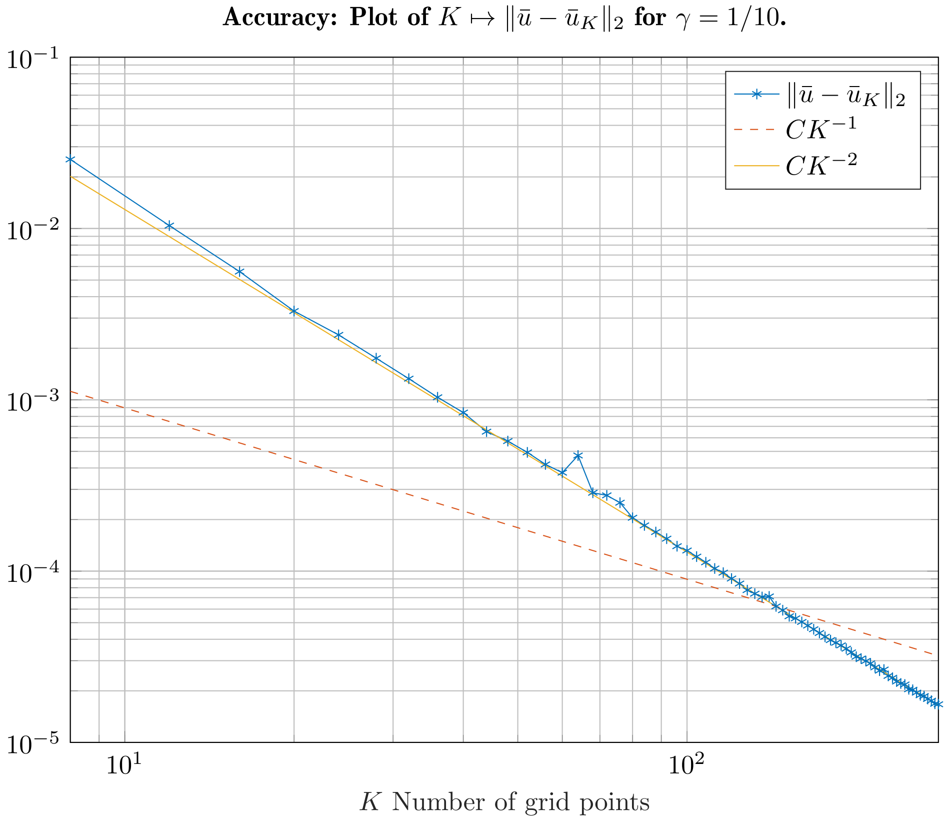
In order to support the claim of quadratic accuracy, we have solved the above constructed problem for every number of grid points $K=1,\dots,300$. We calculate the mappings of interest

$$K\mapsto \big(\|\bar{u}-\bar{u}_K\|_2, \|\bar{y}-\bar{y}_K\|_2, \|\bar{q}-\bar{q}_K\|_2\big), \qquad \bar{y}_K := G_K(\bar{u}_K),\ \bar{q}_K := Q_K(\bar{u}_K)$$

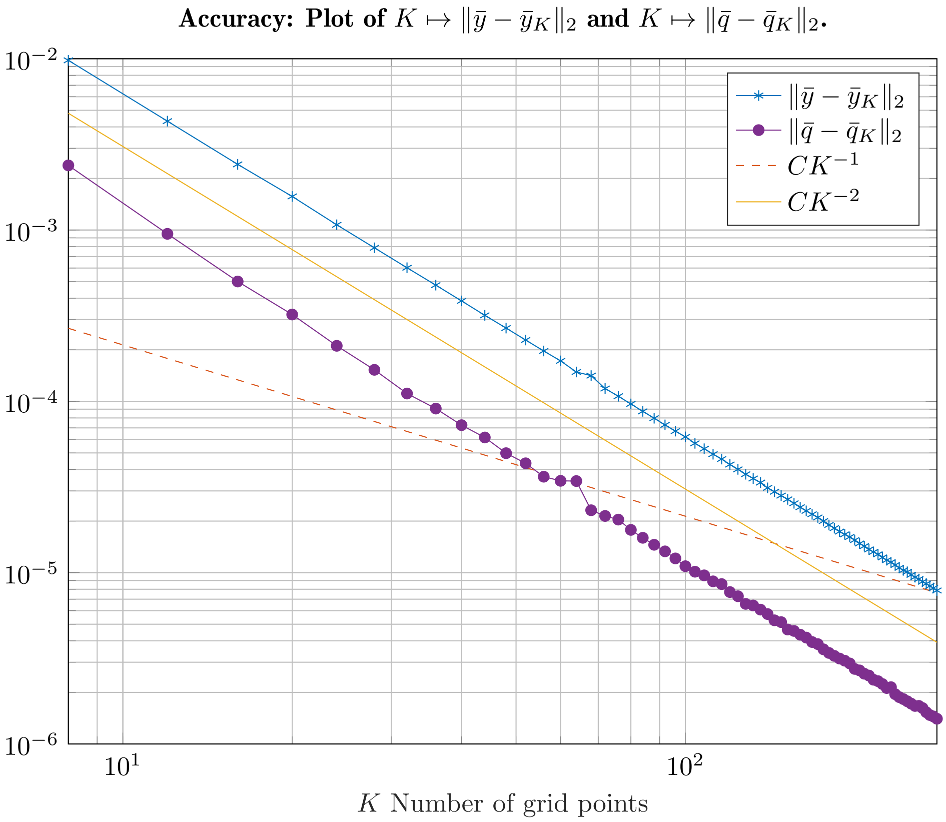
and plot them for $K=5,10,\dots,300$ in a log log–plot; see the figures below. For illustrative purposes, we have included two reference functions $K\mapsto CK^{-1}$ in a dashed line and $K\mapsto CK^{-2}$ in a solid line (for a suitable constant $C>0$). Thus, it is evident that we have second–order convergence for the control, state and adjoint.

Algorithm 1 (NCG scheme)

```
Compute $y_0 = G_K(u_{K,0})$, $q_0 = Q_K(u_{K,0})$ via (6.59)–(6.60).
Compute $d_0 = \hat{J}_K'(u_{K,0})$ via (6.64).
Set $n = 0$, $n_{max} = 1000$, $\varepsilon = 10^{-7}$.
while $n < n_{max}$ do
    Set $u_{K,n+1} = u_{K,n} + \alpha_n d_n$.    ▷ $\alpha_n$ is obtained with a line–search algorithm.
    Compute $y_{K,n+1} = G_K(u_{K,n+1})$, $q_{K,n+1} = Q_K(u_{K,n+1})$ via (6.59)–(6.60).
    if $\|u_{K,n+1} - u_{K,n}\|_2 < \varepsilon$ then
        set $\bar{u}_K := u_{K,n+1} \in \mathcal{P}_K^2$, $\bar{y}_K := y_{K,n+1}$, $\bar{q}_K := q_{K,n+1} \in \mathcal{P}_K^1$ and terminate.
    end if
    Compute $g_{n+1} = \hat{J}_K'(u_{K,n+1})$ via (6.64).
    Set $d_{n+1} = \beta_n d_n - g_{n+1}$.    ▷ $\beta_n$ is a Fletcher–Reeves step size correction.
    Set $n = n + 1$.
end while
```



Accuracy: Plot of $K \mapsto \|\bar{u} - \bar{u}_K\|_2$ for $\gamma = 1/10$.



Let us conclude this chapter with a summary of our findings. First– and second–order accuracy estimates for an optimal control problem governed by an ODE system with bilinear control mechanism were presented. This problem is closely related to a semidiscrete Galerkin approximation of a Fokker–Planck optimal control problem. The accuracy estimates were obtained based on a variational discretization concept combined with a first– and second–order analysis of optimality of the semidiscrete and continuous optimal control problems. Main emphasis was put on the unconstrained case, where the forward and backward problem was discretized by continuous, piecewise linear polynomials. Piecewise quadratic polynomials were used for the discretization of the controls, which lead to a setting where the optimize–then–discretize approach coincides with the discretize–then–optimize one.

In the presence of box constraints on the control, a piecewise constant discretization for the controls was considered and first–order accuracy estimates were obtained in that case. The theoretical claims were supported with numerical evidence.

7

Accuracy estimates for the Fokker–Planck optimal control problem

All people are smart – some before, others after.

Voltaire, 1694 – 1778

In this chapter, linear and quadratic error estimates for the Fokker–Planck optimal control problem from Chapter 4 are derived. This is done by merging the results from Chapters 4–6 together. More precisely, we show that the control problem under investigation satisfies the conditions of the abstract splitting approach stated in Section 5.1. Then, semidiscrete accuracy estimates can be derived by applying the main results of Chapter 4. In the next section, we recall the discretization concepts and precisely formulate the claim of linear and quadratic accuracy estimates.

7.1 Main results: linear and superlinear accuracy

Let us state the optimal control problem and its semi– and fully discretizations. The minimization problem reads

$$\min_{u\in U_{\mathrm{ad}}^T} \hat{J}(u), \tag{7.1}$$

where $\hat{J}(u) := J(G(u), u)$ with control–to–state map $G : L^\infty(0,T) \to W(0,T)$ given in Definition 2.4.1, where we set $m = 1$. The set of admissible controls U_{ad}^T is defined in (2.35) in the same section. The cost functional J is stated in (4.4) with the assumption (J1) from the same section and regularizing norm $Y = L^2(0,T)$. Concerning the Fokker–Planck problem, the assumptions (F1)–(F7) hold, see Chapter 2. In this chapter, we will write $\varrho = \Theta(u)$, given in Definition 4.1.4, for the adjoint of the FP control problem for the reason that the variable q will be used for the adjoint of the ODE control problem.

Next, we recall the semidiscrete minimization problem

$$\min_{u\in U_{\mathrm{ad}}^T} \hat{J}_h(u), \tag{7.2}$$

where $h > 0$ denotes the spatial mesh size, and $N = N(h)$ is the dimension of the corresponding finite element space $\mathcal{P}_\Omega^N$ with basis $\{\psi_i \mid i = 1,\dots,N\}$ given in Section 5.1. We have $\hat{J}_h(u) := J_h(G_h(u),u)$ with the semidiscrete control–to–state map $G_h : L^\infty(0,T) \to H^1(0,T;\mathcal{P}_\Omega^N)$ from (5.26) given by the semidiscrete Galerkin scheme. The semidiscrete cost functional reads for $h > 0$

$$J_h : H^1(0,T;\mathcal{P}_\Omega^N) \times U_{\mathrm{ad}}^T \to \mathbb{R}, \quad J_h(f,u) := \frac{\beta}{2}\|f - p_h^d\|_{L^2(\Omega_T)}^2 + \frac{\alpha}{2}\|f(T) - p_h^T\|_{L^2(\Omega)}^2 + \frac{\gamma}{2}\|u\|_2^2,$$

where $p_h^d := \mathrm{Proj}_{L^2(\Omega)}^N(p^d)$ and $p_h^T := \mathrm{Proj}_{L^2(\Omega)}^N(p^T)$. The projection $\mathrm{Proj}_{L^2(\Omega)}^N$ from $L^2(\Omega)$ to the finite element space $\mathcal{P}_\Omega^N$ is given in Section 5.1. We also introduce the equivalent formulation for $N \in \mathbb{N}$

$$J_N : H^1(0,T)^N \times U_{\mathrm{ad}}^T \to \mathbb{R}, \quad J_N(y,u) := \frac{\beta}{2}\|y - y^d\|_{2,\mathcal{M}}^2 + \frac{\alpha}{2}|y(T) - y^T|_{\mathcal{M}}^2 + \frac{\gamma}{2}\|u\|_2^2.$$

and for $Y_N : L^2(0,T) \to H^1(0,T)^N$ given in (5.25), it holds that

$$J_h(G_h(u),u) = J_N(Y_N(u),u).$$

The mass matrix corresponding to the finite element approximation on Ω is given by $\mathcal{M}_{ij} := \langle\psi_i,\psi_j\rangle_{L^2(\Omega)}$ and the norms $\|\cdot\|_{2,\mathcal{M}}$ and $|\cdot|_{\mathcal{M}}$ are defined below (5.20). As in (5.42)–(5.43), the functions y^d and y^T are the coefficients corresponding to the finite element approximation of p^d and p^T, that is,

$$p_h^d(t,x) = \sum_{i=1}^N y_i^d(t)\,\psi_i(x) \text{ and } \quad p_h^T(x) = \sum_{i=1}^N y_i^T\,\psi_i(x).$$

The initial value from the ODE Cauchy problem for $y = Y_N(u) \in H^1(0,T)^N$ reads

$$y'(t) = (A + u(t)B)y(t), \quad y(0) = y_0 \in \mathbb{R}^N,$$

and the $N \times N$ matrices are given by $A := \mathcal{M}^{-1}\tilde{A}$, $B := \mathcal{M}^{-1}\tilde{B}$ with

$$\begin{aligned}\tilde{A}_{ij} &= -\int_\Omega \big(a\nabla\psi_j(x) - \psi_j(x)c(x)\big)\cdot\nabla\psi_i(x)\,dx,\\ \tilde{B}_{ij} &= -\int_\Omega \psi_j(x)\,M(x)\cdot\nabla\psi_i(x)\,dx.\end{aligned}$$

Similarly, for $q = Q_N(u) \in H^1(0,T)^N$ we have

$$-q'(t) = \beta(y(t) - y^d(t)) + (A + u(t)\,B)^\top q(t), \quad q(T) = \alpha(y(T) - y^T).$$

Thus, the Galerkin approximations P_N and ϱ_N to $p = G(u)$ and $\varrho = \Theta(u)$, respectively, are given by

$$P_N(t,x) = \sum_{i=1}^N y_i(t)\psi_i(x), \quad \varrho_N(t,x) = \sum_{i=1}^N q_i(t)\psi_i(x). \tag{7.3}$$

We recall from Theorem 5.3.3 the convergences of $P_N \to p$ and $\varrho_N \to \varrho$ in $L^2(0,T;H^1(\Omega))$ and $L^\infty(0,T;L^2(\Omega))$ with linear rate, uniformly in u on U_{ad}^T. In that context, we introduce the norms corresponding to the finite element representation of (7.3)

$$|z|_{N,\psi} := \left\|\sum_{i=1}^N z_i\psi_i(\cdot)\right\|_{L^2(\Omega)}, \quad |z|_{N,\nabla\psi} := \left\|\sum_{i=1}^N z_i\nabla\psi_i(\cdot)\right\|_{L^2(\Omega)} \quad \text{for } z \in \mathbb{R}^N. \tag{7.4}$$

Due to the boundedness of P_N and ϱ_N in $C([0,T];H^1(\Omega))$ by Theorem 5.3.3 e), we have that $|P_N|_{N,\nabla\psi}$ and $|\varrho_N|_{N,\nabla\psi}$ are bounded on $[0,T]$, uniformly in N and on U_{ad}^T.
The fully discrete minimization problem reads

$$\min_{u\in U_{\text{ad},K}^T} \hat{J}_{N,K}(u), \tag{7.5}$$

where K denotes the number of uniform grid points on $[0,T]$. The discrete control space $U_{\text{ad},K}^T$ can be either chosen as the box constrained space of piecewise constant functions on that grid or as the box constrained space of continuous, piecewise quadratic polynomials. The fully discrete reduced cost functional $\hat{J}_K : L^2(0,T) \to \mathbb{R}$ is given by

$$\hat{J}_K(u) := \frac{\beta}{2}\int_0^T |y_K(t) - y_K^d(t)|^2\,dt + \frac{\alpha}{2}|y_K(T) - y^T|^2 + \frac{\gamma}{2}\|u\|_2^2,$$

where $y_K^d = \text{Proj}_K^1(y_d)$. It is important to remark that the computation of $\hat{J}_K$ on $U_{\text{ad},K}^T$ is a finite dimensional problem, since the integral over $[0,T]$ becomes a finite sum. The fully discrete state $y_K = G_{N,K}(u)$ and adjoint $q_K = Q_{N,K}(u)$ are given for all $u \in L^2(0,T)$ by the time finite element method

$$\langle y_K', \phi_i\rangle_2 = \langle (A+uB)y_K, \phi_i\rangle_2, \tag{7.6}$$

$$-\langle q_K', \phi_i\rangle_2 = \langle (A+uB)^\top q_K + \beta(y_K - y_{d,K}), \phi_i\rangle_2, \tag{7.7}$$

for all $i = 0,\dots,K$, where $\{\phi_i \mid i = 0,\dots,K\}$ is the basis of $\mathcal{P}_K^1$ given in (6.22). In the case of $u \in U_{\text{ad},K}^T$ both equations reduce to a linear multistep method.
We want to analyze linear error estimates for an optimal control $\bar{u}$ and a minimizer $\bar{u}_{N,K}$ of the fully discrete problem by the splitting idea

$$\|\bar{u} - \bar{u}_{N,K}\|_2 \le \|\bar{u} - \bar{u}_N\|_2 + \|\bar{u}_N - \bar{u}_{N,K}\|_2.$$

The idea is to apply the abstract results from Chapter 5 on the first term and the error estimates from Chapter 6 on the second term. For the latter, we have to uniformly bound the constants appearing in Chapter 6 as N tends to infinity. We only consider the case of piecewise quadratic controls with box constraints from Section 6.8. A look at estimate (6.55) or (6.57) reveals the constants that needs to be controlled is Λ given in Corollary 6.6.4 and $C^{(1)}, C^{(2)}$ from Lemma 6.6.1 and 6.6.2. This issue is addressed next.

Lemma 7.1.1. *Let (F1)–(F7) and (J1)–(J2) from Chapter 4 hold. Let $\bar{u} \in U_{\text{ad}}^T$ satisfy the first– and second–order assumptions (A1)–(A2). Then, the following holds:*

a) The Fokker–Planck optimal control problem (7.1) *and its semidiscrete problem satisfies the conditions (C1)–(C5) from Section 5.2.*

b) The constant Λ given in Corollary 6.6.4 can be chosen independent of N and K: There exists $\Lambda > 0$, $N_0 \in \mathbb{N}$ and a sequence $(K_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ such that the following holds for all $N = N(h) \ge N_0$ and all $K \ge K_N$:

$$\frac{\Lambda}{2}\|\bar{u}_N - \bar{u}_{N,K}\|_2 \le \left(\hat{J}_N'(\bar{u}_{N,K}) - \hat{J}_N'(\bar{u}_N)\right)(\bar{u}_{N,K} - \bar{u}_N).$$

c) $\hat{J}_N'$ is Lipschitz continuous uniformly in N in the sense that there exists $C = C_{\text{ad}}C_{\text{F}*}C_J$ and $N_0 \in \mathbb{N}$ such that

$$\hat{J}_N'(u)v - \hat{J}_N'(w)v \le C(\|u-w\|_2 + h)\|v\|_2$$

for all $N = N(h) \ge N_0$, $u, w \in U_{\text{ad}}^T$, $v \in L^2(0,T)$.

We remark that the necessity of the sequence $(K_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ comes from the fact that for given $N \in \mathbb{N}$, the fully discrete solution of y and q are defined only for sufficiently large K, see (7.6)–(7.7).
Now, we can formulate the main theorem on semidiscrete accuracy estimates for $\bar{u} \in U_{\text{ad}}^T$ satisfying first– and second–order optimality conditions.

Theorem 7.1.2. *(Main theorem on semidiscrete accuracy estimates)*
Let (F1)–(F7) and (J1)–(J2) from Chapter 4 hold. Let $\bar{u} \in U_{\text{ad}}^T$ satisfy the first– and second–order assumptions (A1)–(A2). Then, there exists $C = C_{\text{ad}}C_{\text{F}*}C_J$ and $N_0 \in \mathbb{N}$ such that the following holds for all $N = N(h) \geq N_0$:

a) There exists $\bar{u}_N \in U_{\text{ad}}$ which is a local minimum of the semidiscrete minimization problem (7.2). *Furthermore, for these controls $(\bar{u}_N)_{N\geq N_0}$, conditions* (C1)–(C6) *are satisfied and the following linear accuracy estimate holds*

$$\|\bar{u} - \bar{u}_N\|_2 \leq Ch.$$

b) If additionally the regularity assumptions $G(\bar{u}), \Theta(\bar{u}) \in L^2(0,T;H^3(\Omega)) \cap H^1(0,T;H^2(\Omega))$ hold, then the accuracy estimate in a) is of second–order.

For the fully discrete accuracy estimates, we need the following assumption on the time discretization scheme: For some control $u \in U_{\text{ad}}$, let $y^{N,K} := Y_{N,K}(u)$, $q^{N,K} := Q_{N,K}(u) \in \mathcal{P}_K^1$ and $p := G(u)$, $\varrho := \Theta(u)$. Then, we assume the following linear error estimate for the space–time finite element scheme: Let $p^{N,K}, \varrho^{N,K}$ from $\mathcal{P}_\Omega^N \times \mathcal{P}_K^1$ with

$$p^{N,K}(t,x) := \sum_{i=1}^N p_{ij}\psi_i(x)\phi_j(t), \quad \varrho^{N,K}(t,x) := \sum_{i=1}^N \varrho_{ij}\psi_i(x)\phi_j(t)$$

be the unique solution to

$$\begin{aligned}
&\int_0^T \langle \partial_t p^{N,K}, \psi_i\rangle_{L^2(\Omega)}\phi_j + \mathcal{F}(p^{N,K}, \psi_i)\phi_j \, dt = 0, && i = 1, \dots, N,\ j = 0, \dots, K \\
&\int_0^T \langle \partial_t \varrho^{N,K}, \psi_i\rangle_{L^2(\Omega)}\phi_j + \mathcal{F}(\psi_i, \varrho^{N,K})\phi_j \, dt = \beta \int_0^T \langle p - p^d, \psi_i\rangle_{L^2(\Omega)}\phi_j \, dt, && i = 1, \dots, N,\ j = 0, \dots, K
\end{aligned}$$

with the corresponding initial and terminal condition, respectively. Then, analogously to Theorem 5.3.3, we assume to have linear error estimates in time and space: There exists $C = C_{\text{ad}}C_{\text{F}*}C_J$, $N_0 \in \mathbb{N}$ and a sequence $(K_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ such that for all $u \in U_{\text{ad}}$, $K \geq K_N$ and $N \geq N_0$ it holds that

$$\|p^{N,K} - p\|_{2,H^1} + \|p^{N,K} - p\|_{\infty,2} \leq C(h+k)(\|p\|_{2,H^2} + \|p\|_{H^1,2}), \tag{7.8}$$

$$\|\varrho^{N,K} - \varrho\|_{2,H^1} + \|\varrho^{N,K} - \varrho\|_{\infty,2} \leq C(h+k)(\|\varrho\|_{2,H^2} + \|\varrho\|_{H^1,2}). \tag{7.9}$$

It is reasonably to assume that a linear convergence rate holds given the regularity of p and ϱ due to analogous results for space–time finite element discretizations for similar parabolic problems, cf [35]. Notice that this implies (G1)–(G2) below, stating that the convergences of the time–finite element discretization of the semidiscrete problem from Chapter 6 are uniformly in N on U_{ad}^T, i.e., there exists $C = C_{\text{ad}}C_{\text{F}*}C_J$, $N_0 \in \mathbb{N}$ and a sequence $(K_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ such that for all $u \in U_{\text{ad}}^T$, $N \geq N_0$ and $K \geq K_N$

$$\|\, |y^K - y|_{N,\psi}\|_{L^2(0,T)} \leq Ck\, \|\, |y|_{N,\psi}\|_{H^1(0,T)}, \tag{G1}$$

$$\|\, |q^K - q|_{N,\nabla\psi}\|_{L^2(0,T)} \leq Ck\, \|\, |q|_{N,\nabla\psi}\|_{H^1(0,T)}. \tag{G2}$$

In Chapter 6, we have considered two discrete control space, the box–constrained space of piecewise constant controls, and a variational discretization concept. We will only consider the latter case in the following theorem, since a variational discretization fits well to the bilinear FP optimal control problem. Hence, the fully discrete optimal controls are given by the fully discrete version of (7.21), and therefore, we consider the space of continuous, box–constrained, piecewise quadratic controls

$$U^T_{\mathrm{ad},K} = \left\{u \in C([0,T]) \,:\, u(t) = \min\left\{u^{\max}, \max\{\phi(t)\,\varphi(t), u^{\min}\}\right\}, \quad t \in [0,T],\ \phi, \varphi \in \mathcal{P}^1_K\right\}.$$

Theorem 7.1.3. *(Main theorem on fully discrete accuracy estimates)*
Let (F1)–(F7) and (J1)–(J2) from Chapter 4 hold. Let $\bar{u} \in U^T_{\mathrm{ad}}$ satisfy the first– and second–order assumptions (A1)–(A2). Then, there exists $C = C_{\mathrm{ad}} C_{\mathrm{F}*} C_J$, $N_0 \in \mathbb{N}$ and a sequence $(K_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ such that the following holds for all $N = N(h) \geq N_0$ and all $K \geq K_N$:

a) There exists local minima $\bar{u}_{N,K} \in U^T_{\mathrm{ad},K}$ of the fully discrete minimization problem (7.5) *that converges strongly to $\bar{u}_N$ in $L^2(0,T)$ as $K \to \infty$, where $\bar{u}_N$ is from Theorem 7.1.2 a).*

b) If assumptions (7.8)–(7.9) *or* (G1)–(G2) *hold, then the following linear accuracy estimate holds*

$$\|\bar{u}_N - \bar{u}_{N,K}\|_2 \leq C(h+k).$$

This concludes the results on accuracy estimates of the FP optimal control problem (7.1) with a variational discretization concept for the time dependent controls. In the next section, these statements are proven.

7.2 Proofs

Let us start with the proof of Lemma 7.1.1.

Proof of part a) from Lemma 7.1.1. In Chapter 4, we have performed a first– and second–order analysis of the minimization problem (7.1), and in this process, conditions (C1), (C2.1)–(C2.3) have been verified. Next, we observe that (C3) follows from the convergences of the Galerkin scheme given in Theorem 5.3.3. Subsequently, we can examine (C4), and for that purpose, we aim to apply Lemma 5.2.2. In the proof of Theorem 6.2.1, we have verified that $\hat{J}_h$ is weakly–lower–semicontinuous on $L^2(0,T)$. An application of Lemma 5.2.2 a) yields the candidates $(\bar{u}_N)_{N\in\mathbb{N}}$ for our semidiscrete local minimizers. Next, we see that $\hat{J}$ and $\hat{J}_N$ are of the form from Lemma 5.2.2 c). Thus, given the weak convergence $\bar{u}_N \rightharpoonup \bar{u}$ in $L^2(0,T)$, we need to show that

$$G_N(\bar{u}_N)(T) \to G(\bar{u})(T) \quad \text{strongly in } L^2(\Omega) \quad \text{and} \quad G_N(\bar{u}_N) \to G(\bar{u}) \quad \text{strongly in } L^2(\Omega_T).$$

This follows immediately from the compactness of G on $C([0,T];L^2(\Omega))$, see Section 4.1, and the linear convergence of the Galerkin scheme from Theorem 5.3.3. More precisely, it holds that

$$\|G_N(\bar{u}_N)(T) - G(\bar{u})(T)\|_{L^2(\Omega)} \leq \|G(\bar{u}_N)(T) - G(\bar{u})(T)\|_{L^2(\Omega)} + \|G_N(\bar{u}_N)(T) - G(\bar{u}_N)(T)\|_{L^2(\Omega)},$$

and the latter term can be estimated against $Ch(\|G(\bar{u}_N)\|_{L^2H^2} + \|G(\bar{u}_N)\|_{L^2H^1})$. We remark, that only the $L^2(0,T;H^2(\Omega)) \cap H^1(0,T;L^2(\Omega))$–regularity is necessary for the linear convergence of the Galerkin scheme. Furthermore, notice that $\|G(\bar{u}_N)\|_{L^2H^2}$ and $\|G(\bar{u}_N)\|_{H^1L^2}$ are bounded uniformly in N since $\|\bar{u}_N\|_\infty \leq C_{\mathrm{ad}}$ due to the box–constraints. Consequently, the desired convergences of $G_N(\bar{u}_N)$ to $G(\bar{u})$ follows. Next, due to the H^2–regularity of p^d and p^T and (5.7), we can conclude that

$$\|G_N(\bar{u}_N) - \mathrm{Proj}^N_{L^2(\Omega)}(p^d)\|^2_{L^2(\Omega_T)} \to \|G(u) - p^d\|^2_{L^2(\Omega_T)},$$
$$\|G_N(\bar{u}_N)(T) - \mathrm{Proj}^N_{L^2(\Omega)}(p^T)\|^2_{L^2(\Omega)} \to \|G(u)(T) - p^T\|^2_{L^2(\Omega)}$$

as $N \to \infty$. Thus, we have verified the conditions to apply Lemma 5.2.2 c) and conclude that $(\bar{u}_N)$ are local minimizers of $\hat{J}_N$ which converge strongly to $\bar{u}$ in $L^2(0,T)$. We therefore obtain that condition (C4) holds.

Lastly, let us verify condition (C5) that concerns the behaviour of the second–order derivatives of $\hat{J}$ and $\hat{J}_N$. We recall the derivatives for $u \in U_{\text{ad}}^T, v \in L^2(0,T)$ and $N \in \mathbb{N}$

$$J''(u)(v,v) = \|z\|^2_{L^2(\Omega_T)} + \int_{\Omega_T} (p - p^d) w \, dt \, dx + \int_{\Omega} (p(T) - p^T) w(T) \, dx + \|z(T)\|^2_{L^2(\Omega)} + \gamma \|v\|_2^2, \quad (7.10)$$

$$\hat{J}''_N(u)(v,v) = \|\xi\|^2_{2,\mathcal{M}} + \int_0^T (y(t) - y_d(t))^\top \mathcal{M} \chi(t) \, dt + (y(T) - y^T)^\top \mathcal{M} \chi(T) + |\xi(T)|^2_{\mathcal{M}} + \gamma \|v\|_2^2. \quad (7.11)$$

The $W(0,T)$–functions $z := G'(u)v$, $w := G''(u)(v,v)$ are given in (2.41) and (2.52). The $H^1(0,T)^N$–functions $\xi := Y'_N(u)v$ and $\chi := Y''_N(u)(v,v)$ have been introduced in Lemma 6.1.3. Due to the Galerkin convergence with linear rate from Theorem 5.3.3, it holds for all $u \in U_{\text{ad}}$ that $y = Y_N(u)$ converges to $p = G(u)$ with linear rate, uniformly in u on U_{ad}^T, in the sense that

$$\left\| \sum_{i=1}^N Y_N(u) \, \psi_i - G(u) \right\|_X \leq C_{\text{ad}} C_{\text{F}*} h \to 0, \qquad \text{as } N \to \infty, \quad (7.12)$$

where X is $L^\infty(0,T;L^2(\Omega))$ or $L^2(0,T;H^1(\Omega))$.

We also have to verify the convergence of the same Galerkin scheme for the Fréchet derivatives of G. First, we observe that Theorem 4.1.1 yields higher regularity of $z = G'(u)v$ and $w = G''(u)(v,v)$. Thus, we may apply Theorem 5.3.3 b) to obtain a linear convergence rate of the Galerkin scheme. Consequently, for all $u \in U_{\text{ad}}^T$, $v \in L^2(0,T)$, we deduce for $\xi = Y'_N(u)v$, $\chi = Y''_N(u)(v,v)$ the following convergences

$$\left\| \sum_{i=1}^N \xi_i \, \psi_i - z \right\|_X \leq C_{\text{ad}} C_{\text{F}*} h \|v\|_2 \to 0, \qquad \text{as } N \to 0, \quad (7.13)$$

and

$$\left\| \sum_{i=1}^N \chi_i \, \psi_i - w \right\|_X \leq C_{\text{ad}} C_{\text{F}*} h \|v\|_2^2 \to 0, \qquad \text{as } N \to 0, \quad (7.14)$$

where X is $L^\infty(0,T;L^2(\Omega))$ or $L^2(0,T;H^1(\Omega))$. The convergence in condition (C5) follows now directly from (7.12), (7.13), (7.14) and the definition of y_d, y^T. We will prove this only for

$$\int_0^T y(t)^\top \mathcal{M} \chi(t) \, dt \to \int_{\Omega_T} p(t,x) \, w(t,x) \, dt \, dx,$$

since the other terms can be treated analogously. Recall that $\mathcal{M}_{ij} = \int_\Omega \psi_i(x) \psi_j(x) \, dx$. Let $\| \cdot \|_q =$

$\|\cdot\|_{L^q(\Omega_T)}$ for $q \in \{1,2\}$ and observe that

$$\begin{aligned}
&\left| \int_0^T y(t)^\top \mathcal{M}\chi(t)\,dt - \int_{\Omega_T} p(t,x)\,w(t,x)\,dt\,dx \right| \\
&\quad \le \int_{\Omega_T} \left| \sum_{i,j=1}^N y_i(t)\,\psi_i(x)\,\chi_j(t)\,\psi_j(x) - p(t,x)\,w(t,x) \right| dt\,dx \\
&\quad \le \left\| \sum_{i,j=1}^N y_i\,\psi_i\,\chi_j\,\psi_j - p\sum_{j=1}^N \psi_j\chi_j \right\|_1 + \left\| p\sum_{j=1}^N \psi_j\chi_j - p\,w \right\|_1 \\
&\quad = \left\| \left(\sum_{i=1}^N y_i\,\psi_i\right)\left(\sum_{j=1}^N \psi_j\,\chi_j\right) - p\sum_{j=1}^N \psi_j\chi_j \right\|_1 + \left\| \left(\sum_{j=1}^N \psi_j\chi_j - w\right) p \right\|_1 \\
&\quad \le \left\| \left(\sum_{i=1}^N y_i\,\psi_i\right) - p \right\|_2 \left\| \sum_{j=1}^N \psi_j\chi_j \right\|_2 + \left\| \sum_{j=1}^N \psi_j\chi_j - w \right\|_2 \|p\|_2.
\end{aligned}$$

Since $\left\|\sum_{j=1}^N \psi_j\chi_j\right\|_2$ converges to $\|w\|_2$, it is bounded by some constant $C = C_{\mathrm{ad}}C_{\mathrm{F}*}$. Consequently, an application of (7.14) yields for all small $h > 0$

$$\left| \int_0^T y(t)^\top \mathcal{M}\chi(t)\,dt - \int_{\Omega_T} p(t,x)\,w(t,x)\,dt\,dx \right| \le C_{\mathrm{ad}}C_{\mathrm{F}*}h\|v\|_2^2.$$

Let $\delta > 0$ from condition (C5) be arbitrary but fixed. Hence, there exists $N_0 \in \mathbb{N}$ such that for all $N = N(h) \ge N_0$, it holds that $C_{\mathrm{ad}}C_{\mathrm{F}*}h < \delta$; notice that N_0 takes the role of h_0 from (C5). This proves condition (C5) for one term, and as mentioned above, the proof for the other terms is done analogously. This concludes the proof of part a). □

Since the conditions (C1)–(C5)) are satisfied, we can prove part b) as follows.

Proof of part b). First, we observe that the existence of the semidiscrete minimizer $(\bar{u}_N)_{N\ge N_0}$ of (7.2) and the L^2–convergence to $\bar{u}$ is given by (C4).

Next, we show how condition (C5) implies the local coercivity of $\hat{J}_N''$ around $\bar{u}_N$ if $\bar{u}$ fulfills the second–order assumption (A2). This proves then that Λ from Corollary 6.6.4 is independent of N. For this purpose, recall the critical cone for $\hat{J}$

$$C_{\bar{u}} = \overline{S_{\bar{u}}}^{L^2} \cap \{v \in L^2(0,T) \,:\, \hat{J}'(\bar{u})v = 0\}.$$

For $\tau \ge 0$ and $u \in U_{\mathrm{ad}}^T$, we introduce the following extended cones for $\hat{J}$ and $\hat{J}_N$ at u:

$$\begin{aligned}
E_u^\tau &:= \overline{S_{\bar{u}}}^{L^2} \cap \big\{v \in L^2(0,T) \,:\, |\hat{J}'(u)v| \le \tau\|v\|_2\big\}, \\
E_u^{\tau,N} &:= \overline{S_{\bar{u}}}^{L^2} \cap \big\{v \in L^2(0,T) \,:\, |\hat{J}_N'(u)v| \le \tau\|v\|_2\big\}.
\end{aligned}$$

Obviously, $C_u = E_u^0 \subset E_u^\tau \subset E_u^s$ for $0 \le \tau \le s$. These extended cones are useful since, from Theorem 1.3.6, we obtain local coercivity around $\bar{u}$ in the sense that there exists $\varepsilon, \Lambda, \tau > 0$ such that

$$\hat{J}''(u)(v,v) \ge \frac{\Lambda}{2}\|v\|_2^2, \quad \text{for all } v \in E_{\bar{u}}^\tau,\ u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; L^2). \tag{7.15}$$

Condition (C5) implies that local coercivity holds for $\hat{J}_N$ around $\bar{u}_N$ for all N larger than some $N_0 \in \mathbb{N}$ since (7.15) and $\hat{J}''(u)(v,v) \le \hat{J}_N''(u)(v,v) + \delta\|v\|_2^2$ imply

$$\left(\frac{\Lambda}{2} - \delta\right)\|v\|_2^2 \le \hat{J}_N''(u)(v,v)$$

for all $v \in E^\tau_{\bar{u}}$ and $u \in U_{\mathrm{ad}} \cap B_\varepsilon(\bar{u}; L^2)$. Furthermore, due to the strong L^2–convergence of $\bar{u}_N \to \bar{u}$, we find some $N_1 \in \mathbb{N}$ and $\varepsilon^* > 0$ such that

$$B_{\varepsilon^*}(\bar{u}_N; L^2) \subset B_\varepsilon(\bar{u}; L^2), \quad N \geq N_1. \tag{7.16}$$

Let $\tau > 0$ be given by (7.15) and define $\tau' := \tau/2$. Thus, it is only left to show that there exists some $N_2 \in \mathbb{N}$ such that $E^{\tau',N}_{\bar{u}_N} \subset E^\tau_{\bar{u}}$ for all $N \geq N_2$. For that purpose, let $v^* \in E^{\tau',N}_{\bar{u}_N}$ be arbitrary and observe that for $\bar{y}_N := G_N(\bar{u}_N)$, $\bar{q}_N := Q_N(\bar{u}_N)$

$$|\hat{J}'_N(\bar{u}_N)v^*| = \left|\langle \bar{q}_N^\top \tilde{B}\, \bar{y}_N, v^* \rangle_2 + \langle \bar{u}_N, v^* \rangle_2 \right| \leq \tau' \|v^*\|_2. \tag{7.17}$$

Next, we obtain by Cauchy–Schwarz for $\bar{p} := G(\bar{u})$, $\bar{\varrho} := \Theta(\bar{u})$

$$\begin{aligned} |\hat{J}'(\bar{u})v^*| &= \left|\langle \langle \bar{p}M, \nabla \bar{\varrho} \rangle_{L^2(\Omega)}, v^* \rangle_2 + \langle \bar{u}, v^* \rangle_2 \right| \\ &\leq \|\bar{q}_N^\top \tilde{B} \bar{y}_N - \langle \bar{p}M, \nabla \bar{\varrho} \rangle_{L^2(\Omega)} \|_2 \|v^*\|_2 \\ &\quad + \|\bar{u} - \bar{u}_N\|_2 \|v^*\|_2 + |\langle \bar{q}_N^\top \tilde{B} \bar{y}_N, v^* \rangle_2 + \langle \bar{u}_N, v^* \rangle_2|. \end{aligned}$$

The first two terms tend to zero as N tends to infinity due to the strong L^2–convergence $\bar{u}_N \to \bar{u}$ and $\bar{q}_N^\top \tilde{B}\, \bar{y}_N \to \langle \bar{p}M, \nabla \bar{q} \rangle_{L^2(\Omega)}$ in $L^2(0,T)$ which is shown in the proof of part c) below. Therefore, we find $N_2 \in \mathbb{N}$ such that for all $N \geq N_2$ it holds that

$$\|\bar{q}_N^\top \tilde{B} \bar{y}_N - \langle pb, \nabla q \rangle_{L^2(\Omega)} \|_2 \|v^*\|_2 + \|\bar{u} - \bar{u}_N\|_2 \|v^*\|_2 \leq \tau' \|v^*\|_2.$$

The third term can also be estimated by $\tau' \|v^*\|_2^2$ due to (7.17). Thus, we have shown that

$$|\hat{J}'(\bar{u})v^*| \leq \tau \|v^*\|_2,$$

and hence, $v^* \in E^\tau_{\bar{u}}$. Since v^* was arbitrary, we have proven that $E^{\tau',N}_{\bar{u}_N} \subset E^\tau_{\bar{u}}$ for all $N \geq N_2$. According to (7.15) and (C5) with the choice $\delta = \Lambda/4$, we obtain local coercivity of $\hat{J}''_N$ around $\bar{u}_N$ in the sense that for ε^* from (7.16) and $N^* := \max\{N_0, N_1, N_2\}$, it holds for all $N \geq N^*$ that

$$\hat{J}''_N(u)(v,v) \geq \frac{\Lambda}{4} \|v\|_2^2, \quad v \in E^{\tau',N}_{\bar{u}_N},\ u \in U_{\mathrm{ad}} \cap B_{\varepsilon^*}(\bar{u}_N; L^2). \tag{7.18}$$

This is a sufficient second–order condition on $\bar{u}_N$, that is, $\bar{u}_N$ is a strict local minimum of $\hat{J}_N$ for $N \geq N^*$ that satisfies the standard assumptions for the error analysis, see Theorem 6.8.1 and Theorem 6.8.4. This concludes the proof of part b). □

Next, we prove part c) concerning the Lipschitz continuity of $\hat{J}'_N$, uniformly in N.

Proof of part c). Let $u \in U^T_{\mathrm{ad}}$, $v \in L^2(0,T)$ and recall that

$$\hat{J}'_N(u)v = \beta \langle (Y_N(u) - y^d), \mathcal{M}\, Y'_N(u)v \rangle_2 + \alpha (Y_N(u)_{|t=T} - y^T)^\top \mathcal{M}\, Y'_N(u)v_{|t=T} + \gamma \langle u, v \rangle_2.$$

Analogously to the proof of Lemma 6.2.4, we see that

$$\beta \langle (Y_N(u) - y^d), \mathcal{M}\, Y'_N(u)v \rangle_2 = \langle Q_N(u)\mathcal{M}, vBY_N(u) \rangle_2 + Q_N(u)^\top_{|t=T} \mathcal{M}\, Y'_N(u)v_{|t=T},$$

and hence,

$$\hat{J}'_N(u)v = \langle Q_N(u)^\top \tilde{B}\, Y_N(u), v \rangle_2 + \gamma \langle u, v \rangle_2.$$

Let $q := Q_N(u)$ and $y := Y_N(u)$. It is very important to observe that due to the Galerkin approximation (7.3), we can rewrite the first term for $t \in [0,T]$ as follows

$$\begin{aligned} q(t)^\top \tilde{B}\, y(t) &= \sum_{i,j=1}^{N} \int_\Omega q_i(t)\, \nabla\psi_i(x)^\top\, M(x)\, y_j(t)\psi_j(x)\, dx \\ &= \int_\Omega \nabla\varrho_N(t,x)^\top\, M(x)\, P_N(t,x)\, dx. \end{aligned}$$

Thus, due to the linear convergence rate (uniformly on U_{ad}^T) of

$$\varrho_N \to \Theta(u) \quad \text{in } L^2(0,T;H^1(\Omega)), \qquad P_N \to G(u) \quad \text{in } L^\infty(0,T;L^2(\Omega)),$$

it follows that there exists $C = C_{\text{ad}}C_{\text{F}*}C_J$ such that for all $u \in U_{\text{ad}}^T$ and $N = N(h)$ we have

$$|\langle Q_N(u)^\top \tilde{B}\, Y_N(u), v\rangle_2 - \langle \Theta(u)^\top B\, G(u), v\rangle_2| \le Ch\|v\|_2. \tag{7.19}$$

Furthermore, $\hat{J}'$ is Lipschitz continuous and independent of N, and therefore,

$$|\langle \nabla\Theta(u)^\top M\, G(u), v\rangle_2 - \langle \nabla\Theta(w)^\top M\, G(w), v\rangle_2| \le C\|u-w\|_2\|v\|_2, \tag{7.20}$$

where $C = C_{\text{ad}}C_{\text{F}*}C_J$ is independent of u, w and N. We use the triangle inequality and both estimates (7.19) and (7.20) to conclude

$$\begin{aligned} |\hat{J}_N'(u)v - \hat{J}_N'(w)v| &= \langle Q_N(u)^\top \tilde{B}\, Y_N(u) - Q_N(w)^\top \tilde{B}\, Y_N(w), v\rangle_2 + \gamma\langle u-w, v\rangle_2 \\ &\le C\left(\|u-w\|_2 + h\right)\|v\|_2, \end{aligned}$$

where C is independent of $N = N(h)$ and u, w. This concludes the proof of Lemma 7.1.1. □

Next, we prove the main theorem on semidiscrete accuracy estimates.

Proof of Theorem 7.1.2. In Lemma 7.1.1, we have proven that (C1)–(C5) from Section 4.2 hold. This implies the existence of local minima $\bar{u}_N \in U_{\text{ad}}^T$ of (7.2) that converge strongly in $L^2(0,T)$ to $\bar{u}$. Let us show the uniform convergence of $\bar{u}_N \to \bar{u}$ on $[0,T]$ as $N \to \infty$. For that purpose, recall

$$\Phi[\bar{u}](t) := -\int_\Omega \bar{p}(t,x)\nabla\bar{\varrho}(t,x)^\top M(x)\, dx$$

from (4.11) and let $\Phi_N[\bar{u}_N](t) := \bar{q}_N(t)^\top B\, \bar{y}_N(t)$, where $\bar{q}_N := Q_N(\bar{u}_N)$ and $\bar{y}_N := Y_N(\bar{u}_N)$. This implies

$$\Phi_N[\bar{u}_N](t) = -\int_\Omega \bar{P}_N(t,x)\nabla\bar{\varrho}_N(t,x)^\top M(x)\, dx, \quad t \in [0,T],$$

Due to the FONC of the minimization problem (see Theorem 4.1.5) and the FONC (6.43) of the semidiscrete minimization problem, the following implicit representations hold for $t \in [0,T]$

$$\bar{u}(t) = \min\left\{u^{\max}, \max\left\{-\frac{1}{\gamma}\Phi[\bar{u}](t), u^{\min}\right\}\right\}, \quad \bar{u}_N(t) = \min\left\{u^{\max}, \max\left\{-\frac{1}{\gamma}\Phi_N[\bar{u}_N](t), u^{\min}\right\}\right\}. \tag{7.21}$$

Thus, it is sufficient to prove the uniform convergence of the continuous functions $\Phi_N[\bar{u}_N] \to \Phi[\bar{u}]$ on $[0,T]$. This follows from the convergences of P_N and ϱ_N from Theorem 5.3.3 a) and e).
Next, we have to verify that condition (C6) holds and observe that

$$\hat{J}'(\bar{u}_N)v - \hat{J}_N'(\bar{u}_N)v = \langle \Phi[\bar{u}_N] - \Phi_N[\bar{u}_N], v\rangle_{L^2(0,T)}.$$

Applying the Cauchy-Schwarz inequality, it remains to analyze the convergence rate of

$$\|\Phi[\bar{u}_N] - \Phi_N[\bar{u}_N]\|_2 = \Big\| \int_\Omega \left(\nabla \bar{\varrho}^\top M \bar{p} - \nabla \bar{\varrho}_N^\top M \bar{p}_N\right) dx \Big\|_2,$$

where $\bar{p} = G(\bar{u}_N)$ and $\bar{\varrho} = \Theta(\bar{u}_N)$. Since the convergence rates from Theorem 5.3.3 are uniform for controls on U_{ad}^T, we have that

$$\|\bar{p} - \bar{P}_N\|_{L^\infty(0,T;L^2(\Omega))} + \|\bar{\varrho} - \bar{\varrho}_N\|_{L^2(0,T;H^1(\Omega))} \leq Ch.$$

Since $M \in L^\infty(\Omega)^d$, a Hölder estimate yields the first claim of (C6) with linear rate $r = 1$.
If $\bar{p}$ and $\bar{\varrho}$ satisfy the higher regularity, then Theorem 5.3.3 d) yields the quadratic rates

$$\|\bar{p} - P_N\|_{L^\infty(0,T;L^2(\Omega))} + \|\bar{\varrho} - \bar{\varrho}_N\|_{L^2(0,T;H^1(\Omega))} \leq Ch^2,$$

where $C = C_{\text{ad}} C_{\text{F}*} C_J$ depends additionally on $\bar{p}$ and $\bar{\varrho}$ in the $L^2(0,T;H^3(\Omega))$– and $L^\infty(0,T;H^2(\Omega))$–norm.
Lastly, we have to prove that $(\bar{u}_N - \bar{u})$ is in the extended cone of $\hat{J}$ at $\bar{u}$

$$E_{\bar{u}}^\tau = \left\{ v \in \overline{S_{\bar{u}}}^U \ : \ |J'(\bar{u})v| \leq \tau \|v\|_2 \right\}$$

for sufficiently large N. Let $\tau' := \tau/\sqrt{T}$ and observe that it is sufficient to show that for sufficiently large N (depending on τ'), it holds that

$$\bar{u}_N(t) \begin{cases} \geq \bar{u}(t), & \text{if } \bar{u}(t) = u^{\min}, \\ \leq \bar{u}(t), & \text{if } \bar{u}(t) = u^{\max}, \\ = \bar{u}(t), & \text{if } |\Phi[\bar{u}](t) + \gamma \bar{u}(t)| > \tau', \end{cases} \tag{7.22}$$

since this implies

$$\begin{aligned} |\hat{J}'(\bar{u})(\bar{u}_N - \bar{u})| &= |\langle \Phi[\bar{u}] + \gamma \bar{u}, \bar{u}_N - \bar{u} \rangle_{L^2(0,T)}| \\ &\leq \int_{\{t \in [0,T] : |\Phi[\bar{u}](t) + \gamma \bar{u}(t)| < \tau'\}} \left| (\Phi[\bar{u}](t) + \gamma \bar{u}(t))^\top (\bar{u}_N(t) - \bar{u}(t)) \right| dt \\ &\leq \tau' \sqrt{T} \|\bar{u} - \bar{u}_N\|_2 \\ &= \tau \|\bar{u} - \bar{u}_N\|_2. \end{aligned}$$

Due to the uniform convergence of $\Phi_N[\bar{u}_N]$ to $\Phi[\bar{u}]$ and $\bar{u}_N$ to $\bar{u}$ on $[0,T]$ there exists $N_{\tau'} \in \mathbb{N}$ (depending on τ') such that for all $N \geq N_{\tau'}$ and all $t \in [0,T]$ we have

$$|\Phi[\bar{u}](t) + \gamma \bar{u}(t) - \Phi_N[\bar{u}_N](t) - \gamma \bar{u}_N(t)| < \tau'/2. \tag{7.23}$$

Let $t' \in [0,T]$ such that $|\Phi(t') + \gamma \bar{u}(t')| > \tau'$; notice that if such t' does not exist, then (7.22) is (trivially) satisfied. In order to prove (7.22), we exploit the implicit representations for $\bar{u}$ and $\bar{u}_N$ given in (7.21). Thus, it suffices to show $\Phi_N[\bar{u}_N](t') \neq 0$ as this implies $\bar{u}_N(t')$ is equal to $u^{\min}$ or $u^{\max}$. For that purpose, let $N \geq N_{\tau'}$ and assume $\Phi_N[\bar{u}_N](t') > 0$. This yields with (7.23) the contradiction

$$\Phi_N[\bar{u}_N](t') = \Phi_N[\bar{u}_N](t') - \Phi[\bar{u}](t') + \Phi[\bar{u}](t') > \Phi[\bar{u}](t') - \tau'/2 = \tau' - \tau'/2 > 0,$$

and we can argue analogously if $\Phi_N[\bar{u}_N](t') < 0$. This proves $\Phi_N[\bar{u}_N](t') = 0$, and hence $\bar{u}_N - \bar{u} \in E_{\bar{u}}^\tau$. Therefore, the FP minimization problem and its semidiscretization satisfy conditions (C1)–(C6), and we can apply Theorem 5.2.1 b) to obtain the linear and quadratic rates for h. This concludes the proof of the results on the semidiscretization. □

The last proof of this section is about the theorem on fully discrete accuracy estimate.

Proof of Theorem 7.1.3. Assertion a) was shown in Lemma 6.8.2, where the proof that $\bar{u}_{N,K} \in U^T_{\mathrm{ad},K}$ is given in Theorem 6.8.1.

Let us prove claim b). For the accuracy estimate of the time discretization, we follow the lines of (6.57) and obtain

$$\begin{aligned}\frac{\Lambda}{2}\|\bar{u}_N - \bar{u}_{N,K}\|_2^2 &\leq \left(\hat{J}'_N(\bar{u}_N) - \hat{J}'_N(\bar{u}_{N,K})\right)(\bar{u}_N - \bar{u}_{N,K})\\ &\leq \left(\hat{J}'_{N,K}(\bar{u}_{N,K}) - \hat{J}'_N(\bar{u}_{N,K})\right)(\bar{u}_N - \bar{u}_{N,K}),\end{aligned}$$

where $\Lambda > 0$ is given in Lemma 7.1.1 a) and independent of N and K.

Next, we have to derive a linear convergence rate in k of $\hat{J}'_{N,K}(\bar{u}_{N,K}) - \hat{J}'_N(\bar{u}_{N,K})$ under the assumption of (G1)–(G2).

Let $N_0 \in \mathbb{N}$ be given by Theorem 7.1.2, $N \geq N_0$ be arbitrary and let $K_N \in \mathbb{N}$ (possibly depending on N) be sufficiently large such that for all $K \geq K_N$

$$y^N := Y_N(\bar{u}_{N,K}), \quad q^N := Q_N(\bar{u}_{N,K}), \quad y^{N,K} := Y_{N,K}(\bar{u}_{N,K}), \quad q^{N,K} := Q_{N,K}(\bar{u}_{N,K})$$

exists, see (7.6)–(7.7). Then, for all $v \in L^2(0,T)$ we have

$$\left(\hat{J}'_{N,K}(\bar{u}_{N,K}) - \hat{J}'_N(\bar{u}_{N,K})\right) v = \int_0^T \left(q^{N,K}(t)^\top \tilde{B}\, y^{N,K}(t) - q^N(t)^\top \tilde{B}\, y^N(t)\right) v(t)\, dt.$$

After applying the Cauchy–Schwarz inequality, inserting the mixed term $\pm {q^N}^\top \tilde{B}\, y^{N,K}$ and using the triangle inequality, we need to establish accuracy estimates for terms

$$\left\|\left(q^{N,K} - q^N\right)^\top \tilde{B}\, y^{N,K}\right\|_{L^2(0,T)} \quad \text{and} \quad \left\|{q^N}^\top \tilde{B}\, \left(y^{N,K} - y^N\right)\right\|_{L^2(0,T)}. \tag{7.24}$$

By the definition of $\tilde{B}$, we obtain for the first term

$$\begin{aligned}\left\|\left(q^{N,K} - q^N\right)^\top \tilde{B}\, y^{N,K}\right\|_{L^2(0,T)} &= \left\|\sum_{i,j=1}^N \int_\Omega \left(q_i^{N,K} - q_i^N\right) \nabla\psi_i(x)^\top M(x)\, y_j^{N,K} \psi_j(x)\, dx\right\|_{L^2(0,T)}\\ &\leq \left\|\left\|\sum_{i=1}^N \left(q_i^{N,K} - q_i^N\right)\psi_i\right\|_{L^2(\Omega)} \|M\|_{L^\infty(\Omega)} \left\|\sum_{j=1}^N y_j^{N,K}\psi_j\right\|_{L^2(\Omega)}\right\|_{L^2(0,T)}.\end{aligned}$$

Next, applying a Hölder estimate for the $\|\cdot\|_{L^2(0,T)}$–norm and using the notation (7.4), we obtain

$$\begin{aligned}&\|\left(q^{N,K} - q^N\right)^\top \tilde{B}\, y^{N,K}\|_{L^2(0,T)}\\ &\quad\leq \left\|\sum_{i=1}^N \left(q_i^{N,K} - q_i^N\right)\nabla\psi_i\right\|_{L^2(0,T;L^2(\Omega))} \|M\|_{L^\infty(\Omega)} \left\|\sum_{j=1}^N y_j^{N,K}\psi_j\right\|_{L^\infty(0,T;L^2(\Omega))}\\ &\quad= \left\||q^{N,K} - q^N|_{N,\nabla\psi}\right\|_{L^2(0,T)} \|M\|_{L^\infty(\Omega)} \left\||y^{N,K}|_{N,\psi}\right\|_{L^\infty(0,T)}.\end{aligned}$$

The last term of the third line $\left\||y^{N,K}|_{N,\psi}\right\|_{L^\infty(0,T)}$ is bounded by some constant $C_{\mathrm{ad}}C_{\mathrm{F}*}$ uniformly in N, K since it converges to $\|G(\bar{u})\|_{L^\infty(0,T;L^2(\Omega))}$ as N, K tends to infinity, where $K \geq K_N$ (with K_N possibly depending on N), since

$$\left\|\sum_{j=1}^N y_j^{N,K}\psi_j - G(\bar{u})\right\|_{L^\infty L^2} \leq \left\|\sum_{j=1}^N (y_j^{N,K} - y^N)\psi_j\right\|_{L^\infty L^2} + \left\|\sum_{j=1}^N y_j^N\psi_j - G(\bar{u})\right\|_{L^\infty L^2}.$$

The first term of the last line above can be estimated against

$$\| \, |q^{N,K} - q^N|_{N,\nabla\psi} \|_{L^2(0,T)} \leq Ck \, \| \, |q^N|_{N,\nabla\psi} \|_{L^2(0,T)}$$

due to (G2). As N tends to infinity, the term $\| \, |q^N|_{N,\nabla\psi} \|_{L^2(0,T)}$ converges to

$$\|\nabla\Theta(\bar{u})\|_{L^2(0,T;L^2(\Omega))} \leq \|\Theta(\bar{u})\|_{L^2(0,T;H^1(\Omega))} \leq C_{\mathrm{ad}} C_{\mathrm{F}*} C_J.$$

Therefore, it is also bounded uniformly in N, and we obtain

$$\| \, |q^{N,K} - q^N|_{N,\nabla\psi} \|_{L^2(0,T)} \leq C_{\mathrm{ad}} C_{\mathrm{F}*} C_J k.$$

The second term in (7.24) is estimated with (G1) analogously. This concludes the proof of Theorem 7.1.3. □

8 Appendix: Algorithms

In this chapter, we provide instructions for the MATLAB files that have been used in Section 6.9, Algorithm 1. The numerical scheme solves an ODE–constrained optimal control problem with a non–linear conjugate gradient scheme. Furthermore, a test file is implemented that allows to validate the result about second–order accuracy. For comparison, the optimal control problem is also solved with an Euler discretization for state and adjoint problem. The files can be found in the zip file OPC.zip. In order to run the main program, execute the file OPCbilUQ.m; for the test case, run the file OPCbilUQTest.m and for a solver using an Euler discretization scheme, run OPCbilEuler.m. For convenience of the reader, we state Algorithm 1 again below

```
Compute $y_0 = G_K(u_{K,0})$, $q_0 = Q_K(u_{K,0})$ via (6.59)–(6.60).
Compute $d_0 = \hat{J}'_K(u_{K,0})$ via (6.64).
Set $n = 0$, $n_{max} = 1000$, $\varepsilon = 10^{-7}$.
while $n < n_{max}$ do
    Set $u_{K,n+1} = u_{K,n} + \alpha_n d_n$.    ▷ $\alpha_n$ is obtained with a line–search algorithm.
    Compute $y_{K,n+1} = G_K(u_{K,n+1})$, $q_{K,n+1} = Q_K(u_{K,n+1})$ via (6.59)–(6.60).
    if $\|u_{K,n+1} - u_{K,n}\|_2 < \varepsilon$ then
        set $\bar{u}_K := u_{K,n+1} \in \mathcal{P}^2_K$, $\bar{y}_K := y_{K,n+1}$, $\bar{q}_K := q_{K,n+1} \in \mathcal{P}^1_K$ and terminate.
    end if
    Compute $g_{n+1} = \hat{J}'_K(u_{K,n+1})$ via (6.64).
    Set $d_{n+1} = \beta_n d_n - g_{n+1}$.    ▷ $\beta_n$ is a Fletcher–Reeves step size correction.
    Set $n = n + 1$.
end while
```

Next, we describe with briefly the content of the other files. In stateEq.m and adjointEq.m, the equations for the state and adjoint are contained. The desired state y_d is saved in targetTrajectory.m. The functions solveStateEq.m, solveAdjointEq.m solve the state and adjoint problem with an Euler discretization, while the second–order finite element discretization is used in the file solveStateFEM.m and solveAdjointFEM.m, see line 1 of the algorithm. The cost functional is saved in J.m and the computation of the reduced gradients of J is performed in the functions gradHhat2.m and gradJhat2mid.m, see line 2. The optimal control problem is solved with a non–linear conjugate gradient scheme, which is computed in projectedCG.m and linesearch.m, see line 5 and line 10–11. Lastly, proj.m is a projection function that maps a given function f to $\min\{a, \max\{f, b\}\}$ for given box–constraints a, b.

Declaration of originality

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Würzburg, June 3, 2024

Name ....

Bibliography

[1] R. A. Adams and J. J. Fournier, *Sobolev spaces*, Elsevier, 2003.

[2] W. Alt, *On the approximation of infinite optimization problems with an application to optimal control problems*, Applied Mathematics and Optimization, 12 (1984), pp. 15–27.

[3] W. Alt, U. Felgenhauer, and M. Seydenschwanz, *Euler discretization for a class of nonlinear optimal control problems with control appearing linearly*, Computational Optimization and Applications, 69 (2018), p. 825–856.

[4] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Computational Optimization and Applications, 23 (2002), pp. 201–229.

[5] M. S. Aronna and F. Tröltzsch, *First and second order optimality conditions for the control of Fokker-Planck equations*, ESAIM: Control, Optimisation and Calculus of Variations, 27 (2021), p. 15.

[6] J. Bartsch, A. Borzì, F. Fanelli, and S. Roy, *A theoretical investigation of Brockett's ensemble optimal control problems*, Calculus of Variations and Partial Differential Equations, 58 (2019), p. 1–34.

[7] ——, *A numerical investigation of Brockett's ensemble optimal control problems*, Numerische Mathematik, 149 (2021), p. 1–42.

[8] M. Benning, E. Celledoni, M. J. Ehrhardt, B. Owren, and C.-B. Schönlieb, *Deep learning as optimal control problems: Models and numerical methods*, arXiv preprint arXiv:1904.05657, (2019).

[9] A. Borzì, *Modelling with Ordinary Differential Equations: a Comprehensive Approach*, CRC Press, 2020.

[10] A. Borzì, G. Ciaramella, and M. Sprengel, *Formulation and numerical solution of quantum control problems*, SIAM, 2017.

[11] A. Borzì and S. González Andrade, *Second-order approximation and fast multigrid solution of parabolic bilinear optimization problems*, Advances in Computational Mathematics, 41 (2015), pp. 457–488.

[12] A. Borzì, E.-J. Park, and M. V. Lass, *Multigrid optimization methods for the optimal control of convection–diffusion problems with bilinear control*, Journal of Optimization Theory and Applications, 168 (2016), pp. 510–533.

[13] T. Breiten, K. Kunisch, and L. Pfeiffer, *Control strategies for the Fokker-Planck equation*, ESAIM: Control, Optimisation and Calculus of Variations, 24 (2018), pp. 741–763.

[14] T. Breitenbach and A. Borzi, *A sequential quadratic Hamiltonian method for solving parabolic optimal control problems with discontinuous cost functionals*, Journal of Dynamical and Control Systems, 25 (2019), pp. 403–435.

[15] S. C. Brenner, L. R. Scott, and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, vol. 3, Springer, 2008.

[16] W. Brockett, *Minimum attention control*, in Proceedings of the 36th IEEE Conference on Decision and Control, vol. 3, IEEE, 1997, p. 2628–2632.

[17] R. Brown, *XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies*, The Philosophical Magazine, 4 (1828), p. 161–173.

[18] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, John Wiley & Sons, 2016.

[19] J.-B. Caillau, R. Ferretti, E. Trélat, and H. Zidani, *Numerics for finite-dimensional optimal control problems*, hal-03707475, (2022).

[20] E. Casas, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*, Advances in Computational Mathematics, 26 (2007), pp. 137–153.

[21] E. Casas and F. Tröltzsch, *Second order analysis for optimal control problems: improving results expected from abstract theory*, SIAM Journal on Optimization, 22 (2012), p. 261–279.

[22] ——, *Second-order optimality conditions for weak and strong local solutions of parabolic optimal control problems*, Vietnam Journal of Mathematics, 44 (2016), p. 181–202.

[23] W. Choi and Y.-P. Choi, *A sharp error analysis for the DG method of optimal control problems*, AIMS Mathematics, 7 (2022), p. 9117–9155.

[24] K. Chrysafinos and L. S. Hou, *Error estimates for semidiscrete finite element approximations of linear and semilinear parabolic equations under minimal regularity assumptions*, SIAM journal on numerical analysis, 40 (2002), pp. 282–306.

[25] G. Ciaramella, A. Borzì, G. Dirr, and D. Wachsmuth, *Newton methods for the optimal control of closed quantum spin systems*, SIAM Journal on Scientific Computing, 37 (2015), pp. A319–A346.

[26] E. Di Nezza, G. Palatucci, and E. Valdinoci, *Hitchhiker's guide to the fractional Sobolev spaces*, Bulletin des sciences mathématiques, 136 (2012), pp. 521–573.

[27] A. Dontchev and W. Hager, *The Euler approximation in state constrained optimal control*, Mathematics of Computation, 70 (2001), p. 173–203.

[28] A. L. Dontchev, *An a priori estimate for discrete approximations in nonlinear optimal control*, SIAM journal on control and optimization, 34 (1996), pp. 1315–1328.

[29] A. L. Dontchev, W. W. Hager, and V. M. Veliov, *Second-order Runge–Kutta approximations in control constrained optimal control*, SIAM journal on numerical analysis, 38 (2000), p. 202–226.

[30] J. Douglas, Jr and T. Dupont, *Galerkin methods for parabolic equations*, SIAM Journal on Numerical Analysis, 7 (1970), pp. 575–626.

[31] A. Einstein, *Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen*, Annalen der Physik, 4 (1905).

[32] ——, *Investigations on the Theory of the Brownian Movement*, Courier Corporation, 1956.

[33] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, vol. 159, Springer, 2004.

[34] L. C. Evans, *Partial Differential Equations*, vol. 19, American Mathematical Soc., 2010.

[35] G. Fix and N. Nassif, *On finite element approximations to time-dependent problems*, Numerische Mathematik, 19 (1972), pp. 127–135.

[36] C. A. Fletcher and C. Fletcher, *Computational Galerkin Methods*, Springer, 1984.

[37] B. E. Fristedt and L. F. Gray, *A modern approach to probability theory*, Springer Science & Business Media, 2013.

[38] P. Grisvard, *Elliptic problems in nonsmooth domains*, SIAM, 2011.

[39] W. W. Hager, *Runge-Kutta discretizations of optimal control problems*, in System Theory: Modeling, Analysis and Control, Springer, 2000, pp. 233–244.

[40] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, vol. 3, Springer, 2008.

[41] M. Hermann and A. Borzì, *Analysis of discretization of a modified Crank-Nicolson scheme for quantum optimal control problems*, Master thesis, University of Würzburg, (2016).

[42] M. Hinze, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, Computational Optimization and Applications, 30 (2005), pp. 45–61.

[43] S. Hofmann and A. Borzì, *A sequential quadratic hamiltonian algorithm for training explicit RK neural networks*, Journal of Computational and Applied Mathematics, 405 (2022), p. 113943.

[44] F. Hoppe and I. Neitzel, *Purely time-dependent optimal control of quasilinear parabolic PDEs with sparsity enforcing penalization.*, ESAIM: Control, Optimisation & Calculus of Variations, 28 (2022).

[45] A. Kröner and B. Vexler, *A priori error estimates for elliptic optimal control problems with a bilinear state equation*, Journal of computational and applied mathematics, 230 (2009), p. 781–802.

[46] J. Körner and A. Borzì, *Second–order analysis of Fokker–Planck ensemble optimal control problems*, ESAIM: Control, Optimisation and Calculus of Variations, (2022).

[47] ——, *Accuracy estimates for bilinear optimal control problems governed by ordinary differential equations*, Numerical Functional Analysis and Optimization, 44 (2023), p. 564–602.

[48] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang, *Space-time finite element discretization of parabolic optimal control problems with energy regularization*, SIAM Journal on Numerical Analysis, 59 (2021), p. 675–695.

[49] E. H. Lieb and M. Loss, *Analysis*, vol. 14, American Mathematical Soc., 2001.

[50] P.-L. Lions and A.-S. Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Communications on pure and applied Mathematics, 37 (1984), pp. 511–537.

[51] J. d. Los Reyes, P. Merino, F. Rehberg, and F. Tröltzsch, *Optimality conditions for state-constrained PDE control problems with time-dependent controls*, Control and Cybernetics, 37 (2008), p. 5–38.

[52] X. Mao, *Stochastic Differential Equations and Applications*, Elsevier, 2007.

[53] H. Maurer and N. Osmolovskii, *Equivalence of second-order optimality conditions for bang-bang control problems*, Control & Cybernetics, 34 (2005), pp. 927–950.

[54] E. Nelson, *Dynamical Theories of Brownian Motion*, vol. 106, Princeton University press, 2020.

[55] S. Nowak, *$H^{s,p}$ regularity theory for a class of nonlocal elliptic equations*, Nonlinear Analysis, 195 (2020), p. 111730.

[56] P. M. Pardalos and V. A. Yatsenko, *Optimization and control of bilinear systems: theory, algorithms, and applications*, vol. 11, Springer Science & Business Media, 2010.

[57] J. Prüss, *Maximal regularity for abstract parabolic problems with inhomogeneous boundary data in L_p-spaces*, Masaryk University, 2002.

[58] A. Rösch and D. Wachsmuth, *Numerical verification of optimality conditions*, SIAM journal on control and optimization, 47 (2008), pp. 2557–2581.

[59] S. Roy, M. Annunziato, A. Borzì, and C. Klingenberg, *A Fokker-Planck approach to control collective motion*, Computational Optimization and Applications, 69 (2018), pp. 423–459.

[60] C. Schneider and G. Wachsmuth, *Regularization and discretization error estimates for optimal control of ODEs with group sparsity*, ESAIM: Control, Optimisation and Calculus of Variations, 24 (2018), p. 811–834.

[61] E. Süli and D. F. Mayers, *An introduction to Numerical Analysis*, Cambridge university press, 2003.

[62] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, vol. 25, Springer Science & Business Media, 2007.

[63] H. Triebel, *Theory of function spaces. III, volume 100*, Monographs in Mathematics, (2006).

[64] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*, vol. 112, American Mathematical Soc., 2010.

[65] M. Von Smoluchowski, *Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen*, Annalen der Physik, 326 (1906), pp. 756–780.

[66] O. Von Stryk and R. Bulirsch, *Direct and indirect methods for trajectory optimization*, Annals of operations research, 37 (1992), pp. 357–373.

[67] D. Wachsmuth, *The regularity of the positive part of functions in $L^2(I;H^1(\Omega))\cap H^1(I;H^1(\Omega)^*)$ with applications to parabolic equations*, Commentationes Mathematicae Universitatis Carolinae, 57 (2016), p. 327–332.

[68] Z. Wu, J. Yin, and C. Wang, *Elliptic & Parabolic Equations*, World Scientific, 2006.