Theoretical and numerical investigation of optimal control problems governed by kinetic models

A thesis submitted for the degree of
Doctor of natural sciences

Julius-Maximilians-University of Würzburg
Institute for Mathematics

submitted by

Jan Bartsch

Würzburg, November 2021

This thesis was elaborated in the period of time from May 2018 to June 2021 at the Chair of Mathematics IX (Computational Science) of the Julius-Maximilians-University of Würzburg under the supervision of Prof. Dr. Alfio Borzì (University of Würzburg).

Acknowledgements

I would like to thank my supervisor Professor Alfio Borzì for his guidance and support and the possibility to elaborate this thesis. For his great patience with me and his encouraging and inspiring words. Moreover, Francesco Fanelli and Souvik Roy for fruitful collaboration. Thank you for all the discussions and everything I could learn from you.

I would like to thank my colleagues at the Chair of Scientific Computing and the Institute for a very nice and joyful working environment. In particular, Tim for all the talks about professional and personal topics. Special thanks also to Giovanni for helping me understand and implement Monte Carlo methods and working together on related topics and Jacob for reading this thesis and helping me improve it.

Further, I would like to express my gratitude to my colleagues and supervisors of my industrial partner. Thank you for supporting me both in my personal and professional development. With your help, I have learned to look beyond the horizon of my own subject and learned to talk to non-mathematicians and appreciate the advantages and limits of my discipline. I want to thank the CEO for trusting in me and my work, even in uncertainties and the always most friendly way and positive perspective.

I would like to thank Richard Greiner for giving me the opportunity to be his apprentice over several semesters and learn a lot about extraordinary teaching and organisation. Moreover, all the persons of the Servicentrum for their supporting and helpful work.

Last but not least, I would like to thank my friends and family. My time in Würzburg was influenced by a lot of friends with which talks about deep topics were possible as well as having a lot of fun. They were there encouraging me, when I desperately needed it and inspired me with their way of living. Especially, I want to thank Katja Mönius - without you I probably had not continued to study Math.

I want to thank my family, in particular my parents, grandparents and brothers for always supporting me whatever I wanted to do, having an open ear and helping me to find my way.

Affectionate thanks to my admirable girlfriend Laura for cheering me up and being on my side for so long.

Würzburg, June 2021 *Jan Bartsch*

S.D.G.

☧

Abstract

This thesis is devoted to the numerical and theoretical analysis of ensemble optimal control problems governed by kinetic models. The formulation and study of these problems have been put forward in recent years by R.W. Brockett with the motivation that ensemble control may provide a more general and robust control framework for dynamical systems. Following this formulation, a Liouville (or continuity) equation with an unbounded drift function is considered together with a class of cost functionals that include tracking of ensembles of trajectories of dynamical systems and different control costs. Specifically, L^2, H^1 and L^1 control costs are taken into account which leads to non–smooth optimization problems. For the theoretical investigation of the resulting optimal control problems, a well–posedness theory in weighted Sobolev spaces is presented for Liouville and related transport equations. Specifically, existence and uniqueness results for these equations and energy estimates in suitable norms are provided; in particular norms in weighted Sobolev spaces. Then, non–smooth optimal control problems governed by the Liouville equation are formulated with a control mechanism in the drift function. Further, box–constraints on the control are imposed. The control–to–state map is introduced, that associates to any control the unique solution of the corresponding Liouville equation. Important properties of this map are investigated, specifically, that it is well–defined, continuous and Fréchet differentiable. Using the first two properties, the existence of solutions to the optimal control problems is shown. While proving the differentiability, a loss of regularity is encountered, that is natural to hyperbolic equations. This leads to the need of the investigation of the control–to–state map in the topology of weighted Sobolev spaces. Exploiting the Fréchet differentiability, it is possible to characterize solutions to the optimal control problem as solutions to an optimality system. This system consists of the Liouville equation, its optimization adjoint in the form of a transport equation, and a gradient inequality.

Numerical methodologies for solving Liouville and transport equations are presented that are based on a non–smooth Lagrange optimization framework. For this purpose, approximation and solution schemes for such equations are developed and analyzed. For the approximation of the Liouville model and its optimization adjoint, a combination of a Kurganov–Tadmor method, a Runge–Kutta scheme, and a Strang splitting method are discussed. Stability and second–order accuracy of these resulting schemes are proven in the discrete L^1 norm. In addition, conservation of mass and positivity preservation are confirmed for the solution method of the Liouville model. As numerical optimization strategy, an adapted Krylow–Newton method is applied. Since the control is considered to be an element of H^1 and to obey certain box–constraints, a method for calculating a H^1 projection is presented. Since the optimal control problem is non-smooth, a semi-smooth adaption of Newton's method is taken into account. Results of numerical experiments are presented that successfully validate the proposed deterministic framework.

After the discussion of deterministic schemes, the linear space–homogeneous Keilson–Storer master equation is investigated. This equation was originally developed for the modelling of Brownian motion of particles immersed in a fluid and is a representative model of the class of linear Boltzmann equations. The well–posedness of the Keilson–Storer master equation is investigated and energy estimates in different topologies are derived. To solve this equation numerically, Monte Carlo methods are considered. Such methods take advantage of the kinetic formulation of the Liouville equation and directly implement the behaviour of the system of particles under consideration. This includes the probabilistic behaviour of the collisions between particles. Optimal control problems are formulated with an objective that is constituted of certain expected values in velocity space and the L^2 and H^1 costs of the control. The problems are governed by the Keilson–Storer master equation and the control mechanism is considered to be within the collision kernel. The objective of the optimal control of this model is to drive an ensemble of particles to acquire a desired mean velocity and to achieve a desired final velocity configuration. Existence of solutions of the optimal control problem is proven and a Keilson–Storer optimality system characterizing the solution of the proposed optimal control problem is obtained. The optimality system is used to construct a gradient–based optimization strategy in the framework of Monte–Carlo methods. This task requires to accommodate the resulting adjoint Keilson–Storer model in a form that is consistent with the kinetic formulation. For this reason, we derive an adjoint Keilson–Storer collision kernel and an additional source term.

A similar approach is presented in the case of a linear space–inhomogeneous kinetic model with external forces and with Keilson–Storer collision term. In this framework, a control mechanism in the form of an external space–dependent force is investigated. The purpose of this control is to steer the multi–particle system to follow a desired mean velocity and position and to reach a desired final configuration in phase space.

An optimal control problem using the formulation of ensemble controls is stated with an objective that is constituted of expected values in phase space and H^1 costs of the control. For solving the optimal control problems, a gradient–based computational strategy in the framework of Monte Carlo methods is developed. Part of this is the denoising of the distribution functions calculated by Monte Carlo algorithms using methods of the realm of partial differential equations. A standalone `C++` code is presented that implements the developed non–linear conjugated gradient strategy. Results of numerical experiments confirm the ability of the designed probabilistic control framework to operate as desired. An outlook section about optimal control problems governed by non–linear space–inhomogeneous kinetic models completes this thesis.

Zusammenfassung

Diese Arbeit widmet sich der numerischen und theoretischen Analyse von Problemen der optimalen Kontrolle von Ensembles, die durch kinetische Modelle gesteuert werden. Die Formulierung und Untersuchung von Ensemble–Kontrollproblemen wurden in den letzten Jahren von R.W. Brockett vorgeschlagen und vorangetrieben, mit der Motivation, dass Ensemblekontrolle einen allgemeineren und robusteren Rahmen für die Kontrolle von dynamischen Systemen bieten kann. In Anlehnung an diese Formulierung der Ensemble–Steuerung werden eine Liouville– (oder Kontinuitäts–) Gleichung mit unbeschränkter Driftfunktion und eine Klasse von Kostenfunktionalen miteinbezogen, die das Nachverfolgen der Ensembles und verschiedener Kontrollkosten berücksichtigen. Insbesondere werden L^2, H^1 und L^1 Kontrollkosten betrachtet. Für die theoretische Untersuchung der resultierenden Optimalsteuerungsproblemen wird eine Gutgestelltheitstheorie in gewichteten Sobolev–Räumen für die Liouville– und Transportgleichungen vorgestellt. Insbesondere werden Existenz– und Eindeutigkeitsresultate sowie Energieabschätzungen in geeigneten Normen präsentiert; insbesondere in gewichteten Sobolev–Räumen. Dann wird eine Klasse von nicht–glatten Optimalsteuerungsproblemen formuliert mit der Liouville–Gleichung als Nebenbedingung und einem Kontrollmechanismus in der Driftfunktion. Weiterhin werden Box–Einschränkungen angenommen. Die Kontrolle–zu–Zustand Abbildung wird definiert, die zu jeder Kontrolle einen dazugehörigen Zustand als eindeutige Lösung der Liouville Gleichung zuweist. Wichtige Eigenschaften der Kontrolle–zu–Zustand Abbildung werden untersucht; unter anderem deren Wohldefiniertheit, Stetigkeit und Fréchet–Differenzierbarkeit. Die ersten beiden Eigenschaften werden benutzt, um die Existenz von Lösungen zum Optimierungsproblem zu zeigen. Um die Differenzierbarkeit zu beweisen, wird ein Regularitätsverlust erkannt, der bei hyperbolischen Gleichungen zu erwarten ist. Daher muss die Kontrolle–zu–Zustand Abbildung in gewichteten Sobolevräumen untersucht werden. Unter Verwendung der Differenzierbarkeit ist es möglich, die Lösungen des Optimalsteuerungsproblem als Lösungen eines Optimalitätssystem zu charakterisieren. Dieses System besteht aus

der Liouville Gleichung, einer adjungierten Gleichung in Form einer Transportgleichung und einer Gradient Ungleichung.

Weiterhin wird eine numerische Methode zur Lösung der Liouville und Transport–Gleichungen vorgestellt. Zu diesem Zweck werden Approximations– und Lösungsschemata entwickelt und analysiert. Im Speziellen werden für die Approximation des Liouville–Modells und seiner Adjungierten eine Kombination aus einer Kurganov–Tadmor–Methode, einem Runge–Kutta–Schema und einem Strang–Splitting Verfahren diskutiert. Für die zwei entstehenden Verfahren zum Lösen der Gleichungen wird Stabilität und Genauigkeit zweiter Ordnung bezüglich der diskreten L^1–Norm bewiesen. Zusätzlich werden die Wahrscheinlichkeitserhaltung und der Erhalt der Positivität durch die Lösungsmethode der Liouville Gleichung gezeigt. Das resultierende Optimalitätssystem wird durch ein angepasstes halb–glattes Krylov–Newton–Verfahren. Da angenommen wird, dass die Kontrolle in H^1 liegt und weiterhin Box– Beschränkungen erfüllen muss, wird eine Methode zur Berechnung der H^1 Projektion präsentiert. Ergebnisse numerischer Experimente werden vorgestellt, die das vorgeschlagene deterministische Vorgehen erfolgreich validieren.

Nach der Diskussion der deterministischen Verfahren, werden Optimalsteuerungsprobleme mit der linearen, im Raum homogenen Keilson–Storer Master–Gleichung als Nebenbedingung formuliert. Diese Gleichung wurde ursprünglich als Model für Brownsche Bewegung von Partikeln in Fluiden entwickelt und ist ein repräsentatives Model für lineare Boltzmann Gleichungen. Die Gutgestelltheitstheorie der Keilson–Storer Gleichung wird untersucht und Energieabschätzungen in verschiedenen Normen hergeleitet. Um die Gleichungen numerisch zu lösen, wird auf Monte Carlo Methoden zurückgegriffen. Optimalsteuerungsprobleme werden formuliert mit einem Zielfunktional, das aus bestimmten Erwartungswerten im Geschwindigkeitsraum sowie den Kontrollkosten besteht. Das Ziel der optimalen Steuerung im Kollisionskern dieses Modells ist es, ein Ensemble von Teilchen dazu zu bringen, eine gewünschte mittlere Geschwindigkeit zu erreichen und eine gewünschte Endgeschwindigkeitskonfiguration zu erreichen. Zu diesem Zweck wird ein Keilson–Storer Optimalitätssystem, das die Lösung des vorgeschlagenen optimalen Kontrollproblems charakterisiert, abgeleitet und zur Konstruktion einer gradientenbasierten Strategie im Rahmen von Monte–Carlo–Methoden benutzt. Diese Aufgabe erfordert die Anpassung des resultierenden adjungierten Keilson–Storer–Modells an eine Form, die mit der kinetischen Formulierung konsistent ist.

Darüber hinaus wird ein Monte–Carlo–Framework für die Lösung optimaler Kontrollprobleme vorgestellt, die durch räumlich inhomogene kinetische Modelle mit externen Kräften gesteuert werden. Der Schwerpunkt liegt dabei auf einem linearen kinetischen Modell mit dem Keilson–Storer–Kollisionsterm und einer externen

raumabhängigen Kraft als Kontrollmechanismus. Der Zweck dieser Steuerung ist es, ein Ensemble von Trajektorien von Teilchen so zu beeinflussen, dass deren mittlere Geschwindigkeit und Position einer vorgegebenen Trajektorie folgen und eine gewünschte Endkonfiguration im Phasenraum erreichen. Zu diesem Zweck wird eine gradientenbasierte Optimierungsstrategie im Rahmen von Monte–Carlo–Methoden entwickelt. Ein eigenständiges `C++` Programm wird beschrieben, welches das nicht–lineare konjugierte Gradienten Verfahren implementiert. Ergebnisse von numerischen Experimenten zeigen die Fähigkeit des vorgeschlagenen probabilistischen Ansatzes zum Lösen der Optimierungsprobleme. Ein Abschnitt über Optimalsteuerungsprobleme von nicht–linearen kinetischen Modellen rundet diese Arbeit ab.

Contents

Chapter 1

Introduction

"*Contineat cavitas corpuscula minima motu rapidissimo hinc inde agitata*" [1]. In the kinetic theory of gases, it is assumed that matter consists of spherical particles of very small but finite volume [**18, 79**]. These particles are considered moving in straight lines with constant velocity for a certain period of time before colliding with other particles [**53**].

In the free streaming part, the motion of a single particle can be described using Newton's law of motion

$$\ddot{\xi}(t) = \frac{F(\xi(t), t)}{M}, \tag{1.1}$$

where $\xi(t)$ represents the position of the particle at time t, further F the force on the particle, and M its mass. The second-order ordinary differential equation (1.1) can equivalently be written as a system of first-order differential equations for position ξ and velocity η of a particle as follows

$$\dot{\xi}(t) = \eta, \qquad \dot{\eta}(t) = \frac{F(\xi(t), t)}{M}. \tag{1.2}$$

Upon collision, the particles instantaneously change their velocity according to some collision law. When aspire to simulate the behaviour of multi-particle systems, a high-dimensional problem is encountered. This is due to the fact that the system (1.2) has to be solved for every particle. However, it is possible to model the dynamics of the particles as a partial differential equation. For this purpose, a distribution function $f = f(x, v, t)$ can be defined in the phase space spanned by the position $x \in \Omega \subset \mathbb{R}^d$ and velocity $v \in \mathbb{R}^d$, where $d \geq 1$ and Ω is a bounded convex domain with piecewise smooth boundary. In the statistical framework of kinetic theory, a frequently used model that governs the time evolution of f is the Boltzmann equation

[1] [**18**], Chapter 10, §2. Translated from Latin: The cavity contains very small particles in very fast motion, which move rapidly from this side to the other.

[**25**]. This equation and, more in general, different variants of kinetic models have been widely investigated as they provide the intermediate (mesoscopic) scale in the transition between atomistic and continuous description of gas dynamics; see the recent works [**9, 24, 107**] and the references therein. Furthermore, kinetic models play an important role in applications, ranging from aerodynamics and space propulsion [**83**], to microscale electronic devices and materials [**111**], ionized dilute gases [**134**], and high-temperature plasma [**84**].

While great effort has been put in the modelling and numerical solution of kinetic equations, much less is known concerning the related problems of calibration and control, which can be framed as infinite-dimensional optimization problems with integro-differential constraints. Nevertheless, many present and envisioned applications require the development of control strategies that seem unexplored in those regimes where kinetic models have to be implemented in a kinetic framework; see, e.g., [**68, 98, 101**].

For this reason, we follow a research programme that considers a hierarchy of Boltzmann – like equations and different control mechanisms. Specifically, it is the goal of our work to contribute to this development effort considering optimal control problems governed by the following kinetic model

$$\partial_t f(x,v,t) \,+\, \operatorname{div}(a(x,v,t;u)\, f(x,v,t)) \,=\, C[f](x,v,t). \tag{1.3}$$

In this model, we consider the function $u = u(x,t)$ as the control and $C[f]$ as an integral modelling collisions between particles of the same or different species. In some cases, a control may also be included in $C[f]$. Moreover, the divergence operator div acts on x and v. In a $2d$-dimensional setting, we refer to $a = (a^1, a^2)$, $a^1, a^2 \in \mathbb{R}^d$ as a drift function, which defines the evolution of $\zeta = (\xi, \eta)$ in our controlled model as

$$\dot{\zeta}(t) = a(\zeta(t), t; u). \tag{1.4}$$

The formulation of an optimal control problem governed by (1.3) requires the definition of the purpose of the control and its cost. The objective of the acting control is modelled as the minimisation of a cost functional $J(f,u)$ subject to the constraint given by (1.3) with prescribed initial and boundary conditions. Once the existence of an optimal control is stated, its computation can be based on the related first-order optimality conditions. In the Lagrange framework, these conditions result in an optimality system including the governing model, the related optimization adjoint model, and a gradient equation or inequality; see, e.g., [**28**].

Notice that if we neglect collisions such that $C[f] \equiv 0$ in (1.3), the evolution of f is modelled by the hyperbolic Liouville (or continuity) equation

$$\partial_t f(x,v,t) \,+\, \operatorname{div}(a(x,v,t;u)\, f(x,v,t)) \,=\, 0. \tag{1.5}$$

This equation can be derived considering conservation principles on infinitesimal phase-space volumes [**133**].

Using (1.5), we formulate optimal control problems following the idea of ensemble control that was initially proposed by R.W. Brockett in [**32, 33, 34**]. This formulation is motivated by the intention of designing efficient and robust control strategies for steering ensembles of trajectories of dynamical systems in a desired way. For this purpose, the adequate model governing the evolution of the ensembles expressed in terms of a density is the Liouville equation (1.5). In application, this ensemble may represent the probability density of trajectories of multiple trials of a dynamical system with the initial conditions specified by a distribution function, or the physical density of multiple non-interacting systems (e.g., particles). In both cases, the function that determines the dynamics of these systems appears as the drift coefficient of the Liouville equation. Therefore, the Liouville framework allows to lift the problem of controlling a single trajectory of a finite-dimensional dynamical system to the optimal control problem governed by a partial differential equation (PDE) for a continuum (ensemble) of dynamical systems subject to the same control strategy [**12**].

One of the purposes of the Chapters 2 and 3 is to present a theoretical and numerical optimization framework devoted to ensemble optimal control problems that can involve continuity equations; see, e.g., [**8, 47, 61, 74, 106**] for different classes of these equations. We investigate the existence and uniqueness together with a well-posedness theory of the Liouville equation and the transport equation in different function spaces. We start with the classical theory in L^2 spaces and present its extension for H^m spaces in Theorem 2.2. Then, we extend this theory to certain weighted Sobolev spaces in view of our need of higher integrability resulting from the investigation of solutions to the ensemble optimal control problem. We assume that the density function decays sufficiently fast at infinity and consider a class of drifts that are unbounded but at most linearly increasing at infinity. In this setting, we prove existence and uniqueness of solutions of the Liouville equation in weighted Sobolev spaces in Theorem 2.4.

The challenges of the numerical investigation for the Liouville optimal control problems are manifold. One of these challenges is that we are considering a non-linear control mechanism in the Liouville model where the controls multiply the density function, and this product is subject to spatial differentiation. A further challenge posed by our problems is that the numerical approximation of the Liouville equation must guarantee non-negativity and conservation of probability of the computed density in addition to the required properties of accuracy and stability.

We present a novel formulation and analysis of discretization of the Liouville equation and its optimization adjoint model; the latter is called the adjoint Liouville equation and has the form of a transport equation. For the former, we consider the well-known second-order finite-volume Kurganov-Tadmor (KT) discretization scheme for the spatial flux derivatives that results in a generalized monotone upwind scheme for conservation laws (MUSCL). For the temporal discretization, we use the second-order strong stability preserving Runge-Kutta (SSPRK2) discretization scheme. We show that such schemes possess several important properties that are inherent to exact solutions of the Liouville equation, such as conservation of mass and preservation of positivity. These results appear in Lemma 2.7 and Lemma 2.8. With the help of a discrete stability result stated in Lemma 2.10, we show second-order accuracy of our combined method in Theorem 2.5. For discretizing the adjoint Liouville equation, we choose a second-order Strang operator splitting. This leads to solving an equation having the structure of a Liouville equation and an inhomogeneous linear equation. To solve the former, we are able to apply the same strategy as for the Liouville equation. For the latter one it is possible to derive an exact integration formula. For the resulting splitting scheme, we prove a discrete stability result in Lemma 2.12 and second-order accuracy in Theorem 2.6. By virtue of a suitable test-case, we validate the correctness of our implementation.

In Chapter 3, we illustrate our formulation of ensemble optimal control strategy so as to address many possible requirements in applications of this framework. Moreover, turning to the functional structure of the controls' objectives, we notice that ensemble cost functionals are a much less investigated topic, especially in combination with non-smooth costs of the controls. In addition of Brockett's original formulation, we take not only H^1 control costs into account but also L^2 and L^1 control costs. The presence of L^1 costs and box constraints on the values of the controls require further numerical analysis effort due to the resulting lack of Fréchet differentiability of the resulting optimization problem. To prove existence of optimal controls, we introduce the control-to-state map and investigate its properties in Lemma 3.1 and Theorem 3.1. Specifically, we prove the Lipschitz-continuity and Fréchet differentiability of this map. With this preparation, we are able to conclude existence of solutions to our ensemble control problem in Theorem 3.2. Moreover, we derive a uniqueness result under certain smallness assumptions of the data; see Theorem 3.4.

An important step in solving our ensemble optimal control problems is the derivation of the corresponding first-order optimality conditions that consist of the controlled

Liouville equation, its optimization adjoint, and a variational inequality that we may also call (with some abuse of wording) the optimality condition equation.

The numerical solution of the optimality system proceeds along two main steps that are the numerical approximation of the equations involved and their solution by a numerical optimization scheme. For the latter, we apply a semi-smooth Krylov-Newton method to solve the optimal control problem. Since we have L^1 control costs, we cannot apply any technique from the realm of smooth optimization and also have to take sub-differential calculus into account. Furthermore, due to the fact that we also consider the control to obey some box constraints and to be in H^1, we have to perform a suitable projection within every optimization iteration.

Notice that a specific advantage of Brockett's formulation is that the adjoint equation derived using Lagrange framework does not depend on the solution of the model equation since the functional does only depend linearly on the state. Therefore, the two equations can be solved in parallel. This reasoning applies to all our problems and algorithms.

Starting with Chapter 4, we consider the presence of collision and develop probabilistic methods for solving related optimal control problems.

In Chapter 4, we choose a representative linear collision term with a kernel as proposed by J. Keilson and J.E. Storer [**86**]. This collision kernel was originally proposed for a more realistic kinetic modelling of Brownian motion, and later on it has been successfully used in a range of applications including the estimation of transport coefficients [**17**], laser spectroscopy [**16**], and molecular dynamics simulations [**125**], reorientation of molecules in liquid water [**73**], and quantum transport [**88**]. Further, notice that this term allows to mimic strong and weak collision limits [**121**] and that a microscopic derivation of the Keilson-Storer collision term is possible [**72**]. We remark that again much less is known concerning methods for calibration, control, and optimization of these models, especially in those mesoscopic regimes where probabilistic aspects of the evolution of the particles play an essential role. This is the case in the simulation of rarefied gases with high Knudsen number Kn, where simulation by macroscopic equations suffer inaccuracies; see, e.g., [**22, 50, 95, 96**]. The Knudsen number is the ratio of the mean-free path and the characteristic length of the problem. Furthermore, the mesoscopic setting allows to accommodate the case where the coefficients of the model are prescribed probabilistically by some distribution functions [**55**]. Therefore, although formally the Boltzmann equation is a partial-integro differential equation, methods developed in a deterministic context [**28, 71**] cannot be applied in a truly mesoscopic regime where statistical simulation by Monte Carlo methods is required. Another advantage of Monte Carlo methods is the avoidance of the so-called curse of dimensionality. While the error of deterministic integration techniques depends exponentially on the dimensions, this is not the case for the Monte Carlo techniques. These techniques suffer from inherent noise and slow convergence but their error is independent of the dimension. We have that the solution error is of $\mathcal{O}(1/\sqrt{N})$ where N is the number of samples within the Monte Carlo method, see [**37, 90, 116**].

However, when applying our Lagrange optimization framework to the Keilson-Storer master equation, one immediately recognizes that the resulting adjoint model has a complicated structure that does not belong to the realm of kinetic models, which makes the use of Monte Carlo methods for its solution a challenge.

Our main goal in Chapter 4 and 5 is to develop a Monte Carlo simulation and optimization framework that accommodates our Keilson-Storer optimality system in a way that is consistent with the kinetic description of gases [**13**]. In this framework, a well-known Monte Carlo method in the realm of Boltzmann models is the direct simulation Monte Carlo (DSMC) scheme. Notably, the formulation of the DSMC scheme mimics the derivation of the Boltzmann equation [**43**], and it is one of the most important and frequently used methods for determining the behaviour of dilute gases [**100**]. Its application ranges from vacuum technologies to micro-devices where the Knudsen number is large such that the continuum assumption is no longer valid [**23, 70, 110**]. In the case of a linear kinetic model, it is possible to apply a simplified version of the DSMC technique as discussed in [**68, 101**].

We explore two different kinds of control mechanisms. First, we consider the homogeneous Keilson-Storer master equation and a control within the Keilson-Storer collision kernel that shifts the mean velocity of the particles. Therefore, this mechanism may be interpreted as a change in temperature such that the velocity of the particles is influenced by the resulting temperature gradient [**105**]. In Theorem 4.1, we state the existence of optimal solutions for this case.

Subsequently, we turn to the physical more relevant case of the control mechanism being in the external force. Moreover, we discard the assumptions that the distribution function is homogeneous in space. As a collision model, we again use the Keilson-Storer collision term. In comparison to the homogeneous case, one has to take care of the behaviour of the particles at the boundary and the implementation of a suitable numerical scheme for applying the external force and the change in position due to velocity. In Chapter 4, we elaborate on equations for the moments of the linear kinetic equation with the Keilson-Storer collision term. Further, we investigate the well-posedness of such linear kinetic models and state existence of solutions in Theorem 4.2.

For the optimization strategy, we adapt the well-known non-linear conjugate gradient scheme in the Monte Carlo framework. We implement our scheme in a code called MOCOKI and provide pseudocodes that explain all the details of the implementation. A comparison with well–known deterministic schemes and MOCOKI is performed that show the comparability of the two approaches. Furthermore, numerical experiments are executed that demonstrate the ability of MOCOKI to calculate optimal controls to a given optimization problem.

In Chapter 5, we present preliminary results of our investigation of optimal control problems governed by the Boltzmann equation with the non-linear collision kernel introduced by Boltzmann. That is, we consider binary inter-species collisions. Specifically, we assume fully elastic collisions to ensure the conservation of momentum and energy. Notice that the Boltzmann collision kernel is widely used to describe collisions in rarefied gases. In this case, the adjoint collision kernel is linear and can be written as a collision term that has a gain-loss structure without introducing an additional source term. Further, we consider another control mechanism with a linear drift function in which our control is in the coefficients of this function. We present results of numerical experiments in the seven dimensional phase-space-time domain. The results of numerical experiments successfully validate our framework in physical relevant dimensions.

The results presented in this thesis are partially based on the following publications:

J. Bartsch, A. Borzì, F. Fanelli, and S. Roy, *A theoretical investigation of Brockett's ensemble optimal control problems*, Calc. Var. Partial Differential Equations, 58 (2019), p. Paper No. 162, `https://doi.org/10.1007/s00526-019-1604-2`,

J. Bartsch, G. Nastasi, and A. Borzì, *Optimal control of the Keilson-Storer master equation in a Monte Carlo framework*, J. Comput. Theor. Transp., (2021), `https://doi.org/10.1080/23324309.2021.1896552`,

J. Bartsch and A. Borzì, *MOCOKI: A Monte Carlo approach for optimal control in the force of a linear kinetic model*, Comput. Phys. Commun., (2021), `https://doi.org/10.1016/j.cpc.2021.108030`, and

J. Bartsch, A. Borzì, F. Fanelli, and S. Roy, *A numerical investigation of Brockett's ensemble optimal control problems*, submitted to Numerische Mathematik.

1.1. Notation

The notation used throughout this thesis is given below.

Given a domain $\Omega \subset \mathbb{R}^d$, the symbol $C_c^\infty(\Omega)$ denotes the space of infinitely often differentiable functions with compact support in Ω. Given $m \in \mathbb{N}$, we denote by $C^m(\Omega)$ the space of all m-times continuously differentiable functions defined on Ω, and by $C_b^m(\Omega)$ the subspace of $C^m(\Omega)$ formed by functions which are uniformly bounded together with all their derivatives up to the order m. We equip $C_b^m(\Omega)$ with the $W^{m,\infty}$-norm as follows

$$\|v\|_{C_b^m} := \sum_{|\alpha|\leq m} \|D^\alpha v\|_{L^\infty} .$$

where $\alpha \in \mathbb{N}^d$ is a multi-index, with $|\alpha| \coloneqq \sum_{i=1}^d \alpha_i$ and

$$D^\alpha f \coloneqq \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}},$$

denoting the weak α-th derivative.

For $\varepsilon \in]0,1]$, we denote with $C^{0,\varepsilon}(\Omega)$ the classical Hölder space (Lipschitz space if $\varepsilon = 1$), endowed with the norm

$$\|\Phi\|_{C^{0,\varepsilon}} := \sup_{x\in\Omega} |\Phi(x)| + \sup_{\substack{x,y\in\Omega \\ 0<|x-y|\leq 1}} \frac{|\Phi(x)-\Phi(y)|}{|x-y|^\varepsilon} .$$

In particular, $C^{0,1}(\Omega) \equiv W^{1,\infty}(\Omega)$.

For $m \in \mathbb{N}$ and $1 \leq p \leq +\infty$, we denote with $W^{m,p}(\Omega)$ the usual Sobolev space of L^p-functions with all the derivatives up to the order m in L^p; we also set $H^m(\Omega) := W^{m,2}(\Omega)$. Moreover, we denote $H_T^m \coloneqq H_0^m(0,T)$.

For $1 \leq p < +\infty$, let $W^{-m,p}(\Omega)$ denote the dual space of $W^{m,p}(\Omega)$. For any $p \in [1,+\infty]$, the space $L^p_{loc}(\Omega)$ is the set formed by all functions which belong to $L^p(\Omega_0)$, for any compact subset Ω_0 of Ω.

Furthermore, we make use of the so-called Bochner spaces. Given a Banach space $(X, \|\cdot\|_X)$ and a fixed time $T > 0$, we define for $1 \leq p < \infty$, and a generic representative function $\phi = \phi(x,t)$, the spaces

$$L_T^p(X) := L^p\Big([0,T];X\Big) \qquad \text{with norm} \qquad \|\phi\|_{L_T^p(X)} := \left(\int_0^T \|\phi(\cdot,t)\|_X^p \, dt\right)^{\frac{1}{p}},$$

and

$$L_T^\infty(X) \coloneqq L^\infty([0,T];X) \qquad \text{with norm} \qquad \|\phi\|_{L_T^\infty(X)} := \operatorname{ess}\sup_{t\in[0,T]} \|\phi(\cdot,t)\|_X .$$

Further, for $m \in \mathbb{N}$ and a function $\phi = \phi(t)$, we define

$$C_T^m(X) \coloneqq C^m([0,T];X) \qquad \text{with the norm} \qquad \|\phi\|_{C_T^m(X)} \coloneqq \sum_{i=0}^m \max_{t\in[0,T]} \left\|\frac{\mathrm{d}^i}{\mathrm{d}t^i}\phi(t)\right\|_X .$$

Given a sequence $\Big(\Phi_n\Big)_n$, we use the notation $\Big(\Phi_n\Big)_n \prec X$ meaning that $\Phi_n \in X$ for all $n \in \mathbb{N}$ and that this sequence is uniformly bounded in X. This means there exists some constant $M > 0$ such that $\|\Phi_n\|_X \leq M$ for all $n \in \mathbb{N}$.

Given two Banach spaces X and Y, the space $X \cap Y$, endowed with the norm $\|\cdot\|_{X\cap Y} := \|\cdot\|_X + \|\cdot\|_Y$, is still a Banach space.

For every $p \in [1, +\infty]$, we use the notation $\mathbb{L}^p_T(\mathbb{R}^d) := L^p_T(\mathbb{R}^d) \times L^p_T(\mathbb{R}^d)$. Analogously, $\mathbb{H}^1_T(\mathbb{R}^d) := H^1_T(\mathbb{R}^d) \times H^1_T(\mathbb{R}^d)$. In addition, given two vectors u and v in $\mathbb{R}^d$, we write $u \leq v$ if the inequality is satisfied component by component by the two vectors: namely, $u^i \leq v^i$ for all $1 \leq i \leq d$.

Given two operators A and B, we use the standard symbol $[A, B]$ to denote their commutator: $[A, B] := AB - BA$.

Chapter 2

Liouville equation

The purpose of this chapter is to present a theoretical and numerical investigation of the Liouville equation. This is performed in view of the formulation of optimal control problems governed by this equation. We present a rigorous investigation of a class of Liouville problems with unbounded drift functions.

In Section 2.1, we introduce the Liouville equation and some important properties of it. Further, we formulate our control mechanism and the controlled Liouville equation.

The first step of our analysis, carried out in Section, 2.2 consists in investigating the well-posedness of continuity and transport equations with unbounded drift function, which presents the structure (2.5). We refer to Chapter 3 of [**10**] for the case of bounded drifts, and to the cornerstone paper [**62**] for the case of unbounded drifts having at most linear growth at infinity; see also [**6, 7, 8, 59**] and references therein for recent advances.

However, in order to give full rigorous justification to our formulation of ensemble control problems introduced in Chapter 3, we need to extend (in Section 2.2.2) the classical well-posedness theory to a class of weighted Sobolev spaces H^m_k, see Definition 2.1 below. Roughly speaking, a tempered distribution $f \in H^m$ belongs to H^m_k if f and all its derivatives up to order m belong to the measurable space $\Big(L^2(\mathbb{R}^d), (1+|x|)^k\, dx\Big)$. We point out that existence, uniqueness and regularity properties are derived in this context by standard arguments. The key of the analysis reduces to show suitable *a priori* estimates on the solutions in weighted norms. In passing, we mention that the well-posedness theory in weighted spaces can be adapted to L^p-based spaces, for any $1 \leq p < +\infty$.

In Section 2.3.1, we investigate the approximation of the Liouville equation. We consider the well-known second-order finite-volume Kurganov-Tadmor (KT) discretization scheme for the spatial flux derivatives that results in a generalized monotonic upwind scheme for conservation laws (MUSCL). For the temporal discretization, we use the second-order strong stability preserving Runge-Kutta (SSPRK2) discretization scheme. We prove that our SSPRK2-KT scheme is stable and positive preserving subject to a restriction on the time-step size. Further, we prove that our scheme is second-order convergent in the L^1 norm, see Theorem 2.5. This result is less-known in the context of generic finite-volume schemes for linear conservation laws.

For the adjoint Liouville equation, which corresponds to a transport equation with a source term, we use a second-order Strang time-splitting scheme together with the KT spatial discretization scheme. In Section 2.3.2, we prove stability and second-order accuracy for the resulting combined approximation strategy, see Theorem 2.6.

In Section 2.3.3, we verify our implementation and obtain the desired order of accuracy in the test-cases.

2.1. Preliminaries

The Liouville equation is a hyperbolic-type PDE, which arises in diverse areas of sciences as biology, finances, mechanics, and physics; see e.g., [**41, 54, 57, 67, 109**]. It is often used to model the evolution of density functions representing the probability density of multiple trials of a single evolving system, or the physical density of multiple non-interacting systems. In both cases, the function of the dynamics of the ordinary differential equation (ODE) model appears as the drift coefficient of the Liouville equation.

Given some time $T > 0$, consider a smooth vector field $a(x,t)$ over $\mathbb{R}^d$, where $(x,t) \in \mathbb{R}^d \times [0,T]$. We refer to a as the drift function. It is well-known that, if a scalar function f defined on $\mathbb{R}^d \times [0,T]$ satisfies the Liouville equation

$$\partial_t f(x,t) + \operatorname{div}\Big(a(x,t)\, f(x,t)\Big) = 0, \tag{2.1}$$

with some (say) smooth initial datum $f_{|t=0} = f_0$, then we can represent f by the formula

$$f(x,t) = \frac{1}{\det J\Big(t, \Lambda_t^{-1}(x)\Big)}\, f_0\Big(\Lambda_t^{-1}(x)\Big)\,,$$

where $\Lambda_t(x) = \Lambda(x,t)$ denotes the flow map associated to a, $J(x,t) = \nabla_x \Lambda_t(x)$ is its Jacobian matrix, and $\Lambda_t^{-1}(x)$ means the inverse with respect to the space variable, at t fixed. Notice that (2.1) is (1.3) with $C[f] \equiv 0$, only one phase-space variable is considered, and no control mechanism is introduced up to now. By definition of flow map, Λ verifies the following system of ODEs:

$$\partial_t \Lambda(x,t) = a\Big(\Lambda(x,t), t\Big)\,, \qquad \Lambda(x,0) = x\,, \tag{2.2}$$

Further, we denote with $\mathfrak{F}$ the flux associated to (2.1) as follows

$$\mathfrak{F}(x,t) = a(x,t)\, f(x,t). \tag{2.3}$$

In view of physical considerations, it is generally natural to assume an initial condition f_0 verifying $f_0 \geq 0$, together with the normalization $\int_{\mathbb{R}^d} f_0(x)\, dx = 1$. By equation (2.1), if a is sufficiently smooth and satisfies certain growth conditions (see e.g. [**62**]), it is standard to deduce that

$$f(x,t) \geq 0 \qquad \text{and} \qquad \int_{\mathbb{R}^d} f(x,t)\, dx = \int_{\mathbb{R}^d} f_0(x)\, dx = 1, \qquad t \geq 0. \tag{2.4}$$

The first property can be proved by the vanishing viscosity method and the maximum principle or solving along characteristics; see, e.g., [**67, 76**]. The second property follows from an application of the divergence theorem. Nonetheless, we remark that most of our results do not require the latter two assumptions on f_0.

Next, let us introduce the control mechanism we consider throughout this chapter. Motivated by the fact that frequently used control mechanisms are the linear and bilinear ones, we define the drift function as follows

$$a(x,t;u) = a_0(x,t) + a_1\, u_1(t) + x \circ a_2 u_2(t)\,. \tag{2.5}$$

In (2.5), a_0 is a given smooth vector field, $a_1, a_2 \in \mathbb{R}^d$ are given constants, and $u = (u_1, u_2)$ is the control, which we assume to be smooth for the scope of the present discussion. The control u_1 represents a linear control mechanism and u_2 multiplying the state variable x represents the bilinear control term. Both functions u_1 and u_2 are defined on the time interval $[0,T]$ with values in $\mathbb{R}^d$. The symbol $\circ : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ denotes the Hadamard product of two vectors, this is the multiplication component by component.

Corresponding to the drift function (2.5), we define the controlled Liouville equation

$$\partial_t f(x,t) + \operatorname{div}\Big(a(x,t;u)\, f(x,t)\Big) = 0. \tag{2.6}$$

The Liouville equation offers a convenient framework to accommodate any control mechanism and any possible initial distribution including multi-modal ones. More in detail, let us consider the simple case where $a_0 \equiv 0$, $a_1 = a_2 = 1$ in (2.5) and f_0 is a normal Gaussian uni-modal distribution. Then, the Liouville dynamics can be completely described by the first- and second-moment equations that include the controls u_1 and u_2. To illustrate this fact, consider the following average operator applied to an integrable function ϕ

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

In particular, we have the mean $\bar{m}(t) = \mathbb{E}[x](t)$ and the variance $\bar{\sigma}(t) = \mathbb{E}\Big[\Big(x - \bar{m}(\cdot)\Big)^2\Big](t)$. Then, by taking the average of our controlled dynamical system (2.6), we obtain

$$\begin{aligned} \dot{\bar{m}}(t) &= u_1(t) + \bar{m}(t)\, u_2(t), & \bar{m}(0) &= \bar{m}_0\,, \\ \dot{\bar{\sigma}}(t) &= 2\,\bar{\sigma}(t)\, u_2(t), & \bar{\sigma}(0) &= \bar{\sigma}_0\,. \end{aligned} \tag{2.7}$$

Observe that the control u_1 appears as the main driving force of the mean value of the density, and u_2 determines the evolution of the variance of the density. See [**34**] for more details on this construction. However, the validity of this setting is very limited by the assumptions above whereas the Liouville framework allows to accommodate a more general drift and density functions. We remark that also related to this interpretation is the work in [**39**], in which, the authors deal with a time-optimal control problem in the space of Borel measures.

As a final comment, we point out that for the characterization of the solution to our Liouville optimal control problems, we have to deal with an adjoint Liouville problem. This has the structure of a transport problem, specifically it is given by

$$\partial_t q(x,t) + a\Big(x,t;u\Big) \cdot \nabla q(x,t) = g(x,t)\,, \qquad \text{with} \quad q_{|t=0} = q_0\,, \tag{2.8}$$

where g and q_0 depend on the optimization data. See Chapter 3 for details.

2.2. Theory of Liouville and transport equations with unbounded drifts

In this section, we present results concerning the well-posedness theory of Liouville and transport equations in the class of Sobolev spaces. In view of formula (2.5), we are especially interested in the case when the drift function a may be unbounded, but has at most a linear growth at infinity.

In Section 2.2.1, we review the well-posedness theory in classical H^m spaces, for $m \in \mathbb{N}$. We do not present all the proofs and refer to [**6, 59, 62**] for the details and more general results. Motivated by the study of our ensemble optimal control problem, we extend these results to weighted Sobolev spaces in Section 2.2.2.

2.2.1. Classical theory of Liouville and transport equations

We start our discussion by considering the Liouville equation. Notice that our statements can be repeated in a very similar way with just slight modifications also for the adjoint Liouville problem.

Consider the following Liouville initial-value problem

$$\text{(2.9)}\qquad \begin{cases} \partial_t f(x,t) + \operatorname{div}\Big(a(x,t)\, f(x,t)\Big) = g(x,t) & \text{in} \quad \mathbb{R}^d \times (0,T] \\ f_{|t=0} = f_0 & \text{on} \quad \mathbb{R}^d\,. \end{cases}$$

Whenever attempting to solve equation (2.9), we search for weak solutions. Namely, for all $\phi \in C_c^\infty\Big(\mathbb{R}^d \times [0,T[\Big)$, we want to verify the equality

$$\text{(2.10)}\qquad -\int_0^T\int_{\mathbb{R}^d} f\,\partial_t\phi\,dx\,dt - \int_0^T\int_{\mathbb{R}^d} f\,a\cdot\nabla\phi\,dx\,dt \\ = \int_0^T\int_{\mathbb{R}^d} g\,\phi\,dx\,dt + \int_{\mathbb{R}^d} f_0\,\phi(0)\,dx\,.$$

The theory for this equation is classical, at least in the case of a bounded drift function a. The following well-posedness result is adapted to our needs from Theorem 3.19 in [**10**].

Theorem 2.1. *Let us fix $T > 0$ and $m \in \mathbb{N}$, and let $a \in L^1\Big([0,T]; C_b^{m+1}(\mathbb{R}^d)\Big)$, $f_0 \in H^m(\mathbb{R}^d)$ and $g \in L^1\Big([0,T]; H^m(\mathbb{R}^d)\Big)$. Then there exists a unique weak solution f to (2.9), with $f \in C\Big([0,T]; H^m(\mathbb{R}^d)\Big)$. Moreover, there exists a "universal" constant $C > 0$, independent of f_0, a, g, f and T, such that the following estimate holds true for any $t \in [0,T]$*

$$\|f(t)\|_{H^m} \le C\left(\|f_0\|_{H^m} + \int_0^t \|g(\tau)\|_{H^m}\,d\tau\right)\exp\left(C\int_0^t \|\nabla a(\tau)\|_{C_b^m}\,d\tau\right).$$

Remark 2.1. *In the case $m = 0$, one can replace $\|\nabla a\|_{C_b^0}$ with $\|\operatorname{div} a\|_{L^\infty}$ inside the integral in the exponential term.*

Motivated by the study of our optimal control problem with the control mechanism specified in (2.5), we are interested in the case when a may be unbounded, with at most a linear growth at infinity. For the data and the drift in (2.9), we assume the following for given $m \in \mathbb{N}$

$$\text{(2.11)}\qquad \begin{cases} g \in L^1\Big([0,T]; H^m(\mathbb{R}^d)\Big) \quad \text{and} \quad f_0 \in H^m(\mathbb{R}^d) \\ a \in L^1\Big([0,T]; C^{m+1}(\mathbb{R}^d)\Big) \quad \text{with} \quad \nabla a \in L^1\Big([0,T]; C_b^m(\mathbb{R}^d)\Big)\,. \end{cases}$$

Remark 2.2. *Notice that hypotheses (2.11) imply, that $a(\cdot,t)$ has at most linear growth in space at infinity. Specifically, there exists a constant $C > 0$, such that for almost every $(x,t) \in \mathbb{R}^d \times [0,T]$, one has*

$$|a(x,t)| \le C\,c(t)\,(1+|x|) \qquad \textit{for} \qquad c(t) = \|\nabla a(\cdot,t)\|_{L^\infty} \in L^1\Big([0,T]\Big)\,.$$

The condition of at most linear growth at infinity can be proved to be somehow sharp for well-posedness; see, e.g., [**59, 62**] *and the references therein.*

We use an important statement proved by DiPerna and Lions in [**62**] for the case $m = 0$. However, we give a self-contained presentation of its proof, since parts of its proof are adapted in Section 2.2.2 in the context of weighted Sobolev spaces.

Theorem 2.2. *Let $T > 0$ and $m \in \mathbb{N}$ fixed, and let a, f_0 and g satisfy hypotheses* (2.11)*. Then there exists a unique solution $f \in C\big([0,T]; H^m(\mathbb{R}^d)\big)$ to problem* (2.9)*. Moreover, there exists a "universal" constant $C > 0$, independent of f_0, a, g, f and T, such that the following estimate holds true for any $t \in [0,T]$*

$$\|f(t)\|_{H^m} \,\leq\, C\left(\|f_0\|_{H^m} + \int_0^t \|g(\tau)\|_{H^m}\, d\tau\right)\, \exp\left(C\int_0^t \|\nabla a(\tau)\|_{C^m_b}\, d\tau\right). \tag{2.12}$$

We derive the existence of solutions by an application of Theorem 2.1. The first step is to construct a suitable truncation of the drift function. For this purpose, let us introduce a smooth cut-off function $\chi \in C^\infty_c(\mathbb{R}^d)$ such that χ is radially decreasing, $\chi(x) = 1$ for $|x| \leq 1$ and $\chi(x) = 0$ for $|x| \geq 2$. For all real $M > 0$, we define

$$a_M(x,t) \,:=\, \chi\Big(\frac{x}{M}\Big)\, a(x,t)\,. \tag{2.13}$$

Notice that, by assumptions (2.11), we immediately get that $a_M \in L^1_T(C^{m+1}_b)$ for all $M > 0$. Moreover, in view of Remark 2.2 it holds that

$$\big(\nabla a_M\big)_M \,\prec\, L^1_T(C^m_b)\,, \qquad\qquad \text{with} \qquad \|\nabla a_M\|_{L^1_T(L^\infty)} \,\leq\, C\,, \tag{2.14}$$

for a constant $C > 0$ independent of M. Indeed, denoting by $\mathbb{1}_S$ the characteristic function of a set $S \subset \mathbb{R}^d$ and by $B_\varrho(x)$ the ball in $\mathbb{R}^d$ with center x and radius $\varrho > 0$, we can compute with a generic constant $C > 0$

$$\begin{aligned}\|\nabla a_M\|_{L^\infty} \,&=\, \left\|\frac{1}{M}\,\nabla\chi\Big(\frac{x}{M}\Big)\, a \,+\, \chi\Big(\frac{x}{M}\Big)\,\nabla a\right\|_{L^\infty}\\ &\leq\, C\,\frac{1}{M}\,\left\|a\mathbb{1}_{B_{2M}(0)}\right\|_{L^\infty} \,+\, \|\nabla a\|_{L^\infty} \,\leq\, C\,.\end{aligned}$$

The bounds for higher order derivatives follow by analogue arguments after noticing that we gain a factor $1/M$ in front of a at each order of differentiation.

At this point, for each fixed $M > 0$, we can consider the truncated problem

$$\begin{cases}\partial_t f \,+\, \operatorname{div}(a_M\, f) \,=\, g & \text{in} \quad \mathbb{R}^d \times (0,T]\\ f_{|t=0} \,=\, f_0 & \text{on} \quad \mathbb{R}^d,\end{cases} \tag{2.15}$$

which possesses a unique weak solution $f_M \in C\big([0,T]; H^m(\mathbb{R}^d)\big)$, by virtue of Theorem 2.1. Moreover, each f_M satisfies the energy estimate (2.12), up to replacing a by a_M. Thus, we have

$$\|f_M(t)\|_{H^m} \,\leq\, C\left(\|f_0\|_{H^m} + \int_0^t \|g(\tau)\|_{H^m}\, d\tau\right)\, \exp\left(C\int_0^t \|\nabla a_M(\tau)\|_{C^m_b}\, d\tau\right). \tag{2.16}$$

Thanks to property (2.14), we deduce the uniform bounds

$$\Big(f_M\Big)_M \prec L^\infty\Big([0,T];H^m(\mathbb{R}^d)\Big).$$

As a consequence, we obtain the existence of a $f \in L^\infty_T(H^m)$ such that one has $f_M \overset{*}{\rightharpoonup} f$ in $L^\infty_T(H^m)$ up to the extraction of a subsequence.

Our next goal is to show that f actually solves problem (2.9) in the weak form given by (2.10). For this purpose, we need to pass to the limit for $M \to +\infty$, in the weak formulation of (2.15). For any $\phi \in C^\infty_c\Big(\mathbb{R}^d \times [0,T[\Big)$ we have

$$\begin{aligned}(2.17)\qquad -\int_0^T\int_{\mathbb{R}^d} f_M\,\partial_t\phi\,dx\,dt\,-\int_0^T\int_{\mathbb{R}^d} f_M\,a_M\cdot\nabla\phi\,dx\,dt \\ =\int_0^T\int_{\mathbb{R}^d} g\,\phi\,dx\,dt\,+\int_{\mathbb{R}^d} f_0\,\phi(0)\,dx\,.\end{aligned}$$

It is enough to prove the convergence in the case of minimal regularity. Thus, we restrict to the case $m=0$ in the next argument.

The only term which presents some difficulties is (2.17) is the term $f_M\,a_M$. Its convergence is based on the next lemma, whose proof is elementary and hence omitted.

Lemma 2.1. *For all compact set $K \subset \mathbb{R}^d$, it holds that*

$$\|a_M - a\|_{L^1_T(L^\infty(K))} \longrightarrow 0 \qquad\qquad as \quad M \to +\infty\,.$$

Let now K denote the support in x of ϕ, where ϕ is the test function appearing in (2.17). Thanks to the uniform bounds, to the strong convergence of a_M to a in $L^1_T\Big(L^\infty(K)\Big)$ (given by Lemma 2.1) and the weak-$*$ convergence of f_M to f in $L^\infty_T(L^2)$, we deduce that $\Big(f_M\,a_M\Big)_M$ is uniformly bounded in $L^1_T\Big(L^2(K)\Big)$, and $f_M\,a_M \overset{*}{\rightharpoonup} f\,a$ in that space in the limit when $M \to +\infty$.

In the end, we have proved that the limit function f is a weak solution to (2.9). Thanks to (2.16), the uniform bounds (2.14), and lower semi-continuity of the norm, we also deduce that f verifies the energy estimate (2.12).

It remains to prove uniqueness of solutions and their time regularity. They are both consequences of the next proposition.

Proposition 2.1. *Let $T>0$ and take $m \in \mathbb{N}$. Let $f \in L^\infty_T(H^m)$ be a weak solution to equation* (2.9) *under hypotheses* (2.11).
Then $f \in C\Big([0,T];H^m(\mathbb{R}^d)\Big)$ and it verifies the energy estimate (2.12).

We present the proof of the previous claim in the minimal regularity case, namely for $m=0$. The general case follows by the same token. To start with, let us state a classical lemma whose proof is hence omitted; see, e.g., [**6, 62**] for details.

Let us fix a function $s \in C_c^\infty(\mathbb{R}^d)$, with $s \equiv 1$ for $|x| \leq 1$ and $s \equiv 0$ for $|x| \geq 2$, s radially decreasing and such that $\int_{\mathbb{R}^d} s = 1$. For all $n \in \mathbb{N}$, we then define $s_n(x) := n^d\, s(nx)$. We refer to the family $(s_n)_n$ as a family of standard mollifiers.

Lemma 2.2. *Let $(s_n)_n$ be a family of standard mollifiers, as constructed here above. For all $n \in \mathbb{N}$, define the operator S_n, acting on tempered distributions over $\mathbb{R}_+ \times \mathbb{R}^d$, by the formula*

$$S_n f \;:=\; s_n *_x f\,,$$

where the symbol $*_x$ means that the convolution is taken only with respect to the space variable. For given $f \in L^\infty_T(L^2)$ and $a \in L^1_T(C^1)$ such that $\nabla a \in L^1_T(C_b)$, we set, for all $n \in \mathbb{N}$ and $1 \leq j \leq d$,

$$r_n^j(f) \;:=\; \partial_j\left(\Big[a, S_n\Big]f\right)\,.$$

Then, for all j fixed, we have $\left(r_n^j\right)_n \prec L^1_T(L^2)$. Moreover, for $n \to +\infty$, we have the strong convergence $r_n^j \to 0$ in $L^1_T(L^2)$.

Let us also recall the following standard notation. For X a Banach space and X^* its pre-dual, we denote by $C_w\Big([0,T];X\Big)$ the set of measurable functions $f : [0,T] \to X$ which are continuous with respect to the weak topology. Namely, for any $\phi \in X^*$, the function $t \mapsto (\phi, f(t))_{X^* \times X}$ is continuous over $[0,T]$.

With this preparation, we are now ready to prove Proposition 2.1.

Proof of Proposition 2.1. With the same notations as in Lemma 2.2, we define $f_n := S_n f$. Notice that $(f_n)_n \subset L^\infty_T(L^2)$. Moreover, f_n satisfies the equation

$$\partial_t f_n + \operatorname{div}\Big(a\, f\Big) = g_n + r_n\,, \qquad \text{with} \qquad \Big(f_n\Big)_{|t=0} = S_n f_0\,, \tag{2.18}$$

where we have set $r_n := \operatorname{div}\left(\Big[a, S_n\Big]f\right)$. Notice that one has $\|S_n f_0\|_{L^2} \leq C\, \|f_0\|_{L^2}$ and $\|g_n\|_{L^1_T(L^2)} \leq C\, \|g\|_{L^1_T(L^2)}$. Furthermore, when $n \to +\infty$, we have the strong convergences $g_n \to g$ in $L^1_T(L^2)$ and $S_n f_0 \to f_0$ in L^2. In addition, by Lemma 2.2, we know that $\|r_n\|_{L^1_T(L^2)} \leq C$ and $r_n \to 0$ in $L^1_T(L^2)$.

Next, an inspection of (2.18) shows that $\Big(\partial_t f_n\Big)_n \prec L^1_T(H^{-1}_{\rm loc})$, which in turn gives us the uniform embedding $\Big(f_n\Big)_n \prec C_T(H^{-1}_{\rm loc})$. From this latter property, combined with a density argument and the uniform boundedness of $\Big(f_n\Big)_n$ in $L^\infty_T(L^2)$, we deduce that $\Big(f_n\Big)_n$ is uniformly bounded in $C_w\Big([0,T];L^2(\mathbb{R}^d)\Big)$.

Now, let us take the L^2 scalar product of equation (2.18) by f_n. We get using integration by parts that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\|f_n\|^2_{L^2} + \frac{1}{2}\int \operatorname{div} a\, |f_n|^2\, dx = \int g_n\, f_n\, dx\,, \tag{2.19}$$

which implies that $\|f_n(t)\|_{L^2} \in C\Big([0,T]\Big)$ for all $n \in \mathbb{N}$. Thanks to this property, together with the fact that $f_n \in C_w\Big([0,T];L^2(\mathbb{R}^d)\Big)$, after writing

$$\|f_n(t+h) - f_n(t)\|_{L^2}^2 = \|f_n(t+h)\|_{L^2}^2 - 2\,\big(f_n(t+h), f_n(t)\big)_{L^2\times L^2} + \|f_n(t)\|_{L^2}^2\,,$$

one deduces that f_n belongs to $C_T(L^2)$ for all $n \in \mathbb{N}$, since the right-hand side goes to zero as h tends to zero.

Further, relation (2.19) also yields

$$\begin{aligned}\|f_n(t)\|_{L^2} \leq C\,\exp&\left(C\int_0^t \|\mathrm{div}\,a(\tau)\|_{L^\infty}\,d\tau\right)\times \\ &\times\left(\|S_nf_0\|_{L^2} + \int_0^t \big(\|g_n(\tau)\|_{L^2} + \|r_n(\tau)\|_{L^2}\big)\,d\tau\right) \\ \leq C\,\exp&\left(C\int_0^t \|\mathrm{div}\,a(\tau)\|_{L^\infty}\,d\tau\right)\left(\|f_0\|_{L^2} + \int_0^t \|g(\tau)\|_{L^2}\,d\tau\right),\end{aligned} \tag{2.20}$$

for all $t \in [0,T]$, using Grönwall's lemma and the previous properties on $\Big(S_nf_0\Big)_n$, $\Big(g_n\Big)_n$ and $\Big(r_n\Big)_n$. In view of this energy estimate, we deduce that $\Big(f_n\Big)_n$ is uniformly bounded in $C_T(L^2)$.

By a similar argument, using the fact that $\Big(S_nf_0\Big)_n$, $\Big(g_n\Big)_n$ and $\Big(r_n\Big)_n$ are strongly convergent in the respective functional spaces, we can moreover deduce that $\Big(f_n\Big)_n$ is a Cauchy sequence in $C_T(L^2)$. To prove this, we take $m < n$ and consider the difference $\delta^n_m f := f_n - f_m$. Then, $\delta^n_m f$ fulfils

$$\partial_t\delta^n_m f + \mathrm{div}\left(a\,\delta^n_m f\right) = \delta^n_m g + \delta^n_m r\,, \qquad\qquad \text{with} \quad \delta^n_m f_{|t=0} = \delta^n_m f_0 := f^n_0 - f^m_0\,,$$

where we have defined also $\delta^n_m g := g_n - g_m$ and $\delta^n_m r := r_n - r_m$. To this equation we can also apply the energy estimates, and obtain

$$\|\delta^n_m f\|_{L^\infty_T(L^2)} \leq C\,\exp\left(C\,\|\mathrm{div}\,a\|_{L^1_T(L^\infty)}\right)\left(\|\delta^n_m f_0\|_{L^2} + \|\delta^n_m g\|_{L^1_T(L^2)} + \|\delta^n_m r\|_{L^1_T(L^2)}\right)\,.$$

Since $\Big(S_nf_0\Big)_n$, $\Big(g_n\Big)_n$ and $\Big(r_n\Big)_n$ are strongly convergent in the respective functional spaces, they are Cauchy sequences. This implies that the limit f of the sequence $\Big(f_n\Big)_n$ belongs to $C_T(L^2)$, and the convergence $f_n \to f$ is strong in this space. Finally, passing to the limit in the left-hand side of (2.20) we see that f verifies the energy estimate (2.12). □

Now, stability and uniqueness follow from Corollary 2.1.

Corollary 2.1. *Fix $T > 0$ and $m \in \mathbb{N}$, and let a satisfy the assumptions given in* (2.11)*. For $i = 1,2$, take an initial datum $f^i_0 \in H^m(\mathbb{R}^d)$ and an external source $g^i \in L^1\Big([0,T];H^m(\mathbb{R}^d)\Big)$, and let $f^i \in L^\infty_T(H^m)$ be a corresponding solution to* (2.9)*, whose existence is guaranteed by the previous arguments.*

Then, after defining $\delta f_0 := f_0^1 - f_0^2$, $\delta g := g^1 - g^2$ *and* $\delta f := f^1 - f^2$, *for all* $t \in [0,T]$, *for some constant* $C > 0$ *independent of the data and the respective solutions it holds that*

$$\|\delta f(t)\|_{H^m} \leq C \left(\|\delta f_0\|_{H^m} + \int_0^t \|\delta g(\tau)\|_{H^m} \, d\tau \right) \exp\left(C \int_0^t \|\nabla a(\tau)\|_{C_b^m} \, d\tau \right) .$$

Proof. By taking the difference of the equations satisfied by f^1 and f^2, one deduces that $\delta f \in L_T^\infty(L^2)$ is a weak solution to the initial-value problem

$$\begin{cases} \partial_t \delta f \, + \, \mathrm{div}\Big(a\, \delta f\Big) \; = \; \delta g & \text{in} \qquad \mathbb{R}^d \times (0,T] \\ \delta f_{|t=0} \; = \; \delta f_0 & \text{on} \qquad \mathbb{R}^d. \end{cases}$$

To this problem, we can apply Proposition 2.1 and gain the desired estimate. □

The characterisation of controls in our framework requires the solution of an adjoint Liouville problem, which is given by a linear transport problem (2.8). Therefore, we consider the following transport problem to complete the analysis of the present section

$$\begin{cases} \partial_t q \, + \, a \cdot \nabla q \, + \, \mathfrak{a}\, q \; = \; g & \text{in } \mathbb{R}^d \times (0,T] \\ q_{|t=0} \; = \; q_0 & \text{on } \mathbb{R}^d \, . \end{cases} \tag{2.21}$$

We assume that the data q_0, a and g verify the assumptions in (2.11), where f_0 is replaced by q_0. Moreover, we assume that $\mathfrak{a}$ has the same regularity as $\mathrm{div}\, a$, specifically $\mathfrak{a} \in L^1\Big([0,T]; C_b^m(\mathbb{R}^d)\Big)$.

The weak formulation of (2.21) now reads

$$\begin{aligned} -\int_0^T\!\!\int_{\mathbb{R}^d} q\, \partial_t \phi - \int_0^T\!\!\int_{\mathbb{R}^d} q\, a \cdot \nabla \phi - \int_0^T\!\!\int_{\mathbb{R}^d} q \; \mathrm{div}\, a\, \phi + \int_0^T\!\!\int_{\mathbb{R}^d} q\, \mathfrak{a}\, \phi & \\ = \int_0^T\!\!\int_{\mathbb{R}^d} g\, \phi + \int_{\mathbb{R}^d} q_0\, \phi(0) \, , & \end{aligned} \tag{2.22}$$

for all $\phi \in C_c^\infty\Big(\mathbb{R}^d \times [0,T[\Big)$. For (2.21), we have the following well-posedness result, analogous to Theorem 2.2 for the Liouville equation.

Theorem 2.3. *Fix $T > 0$ and $m \in \mathbb{N}$, and let a, $\mathfrak{a}$, q_0 and g satisfy the assumptions stated above. Then there exists a unique solution $q \in C\Big([0,T]; H^m(\mathbb{R}^d)\Big)$ to equation* (2.21). *Moreover, there exists a "universal" constant $C > 0$, independent of q_0, a, $\mathfrak{a}$, g, q and T, such that the following estimate holds true for any $t \in [0,T]$:*

$$\|q(t)\|_{H^m} \leq C \left(\|q_0\|_{H^m} + \int_0^t \|g(\tau)\|_{H^m} \, d\tau \right) \times \\ \times \exp \left(C \int_0^t \left(\|\nabla a(\tau)\|_{C_b^m} + \|\mathfrak{a}(\tau)\|_{C_b^m} \right) d\tau \right) . \tag{2.23}$$

The proof is analogous to the one given for Theorem 2.1. Notice that while passing to the limit in the weak formulation (2.22), at step n of the regularization procedure, one has to deal with the terms $\operatorname{div} a^n$ and $\mathfrak{a}^n$. One can use Proposition 4.21 and Theorem 4.22 of [**31**] to deduce that both terms converge to $\operatorname{div} a$ and $\mathfrak{a}$, respectively, in $L^1_T\Big(L^\infty(K)\Big)$ for $n \to +\infty$.

2.2.2. Well-posedness theory in weighted spaces

In this section, we extend the previous theory to Sobolev spaces with weights. This analysis is especially important for the investigation of the Liouville control-to-state map and of the Liouville ensemble optimal control problem carried out in the next sections. We only present the case of the Liouville equation. However, the statements that follow hold also for the transport problem, with minor modifications in the proofs.

For the analysis of the Liouville control-to-state map, we need to prove weighted integrability of f, due to the growth of the drift function. For this purpose, we introduce the following definition.

Definition 2.1. *Fix $(m,k) \in \mathbb{N}^2$. We define the space $H^m_k(\mathbb{R}^d)$ in the following way:*

$$H^0_k(\mathbb{R}^d) \;=\; L^2_k(\mathbb{R}^d) \;:=\; \Big\{ f \in L^2(\mathbb{R}^d) \;\Big|\; |x|^k \, f \;\in\; L^2(\mathbb{R}^d) \Big\} \,,$$

and, for $m \geq 1$, we set

$$H^m_k(\mathbb{R}^d) \;:=\; \Big\{ f \in H^m(\mathbb{R}^d) \cap H^{m-1}_k(\mathbb{R}^d) \;\Big|\; |x|^k \, D^\alpha f \;\in\; L^2(\mathbb{R}^d) \quad \forall\, |\alpha| = m \Big\} \,.$$

The space H^m_k is endowed with the following norm

$$\|f\|_{H^m_k} := \sum_{|\alpha| \leq m} \left\| \Big(1 + |x|^k\Big) \, D^\alpha f \right\|_{L^2} \,.$$

Sometimes, given $m \in \mathbb{N}$, we use the notation $\|\nabla^m f\|_{L^2} \;=\; \sum_{|\alpha|=m} \|D^\alpha f\|_{L^2}$, and analogous writing for weighted norms.

Notice that, for all m and k in $\mathbb{N}$, one has $H^m_k \;\subset\; H^m$ and $H^m \;=\; H^m_0$. Furthermore, since we want to avoid too singular behaviours close to 0, we often focus on the special case $m \;\leq\; k$, which is enough for our scopes. In that case, we have a simple characterization of the spaces H^m_k, which will be useful especially in Section 3.2, when studying the control-to-state map related to our optimal control problem.

PROPOSITION 2.2. (1) *Given $k \in \mathbb{N}$, one has $f \in L^2_k$ if and only if $(1+|x|^k)\,f \in L^2$.*

(2) *For $k \in \mathbb{N}\setminus\{0\}$ and $1 \leq m \leq k$, let $f \in H^m \cap H^{m-1}_k$. Then $f \in H^m_k$ if and only if $|x|^k\,f \in H^m$.*
In particular, a tempered distribution f belongs to H^1_1 if and only if both f and $|x|\,f$ belong to H^1; it belongs to H^2_2 if and only if both f and $|x|^2\,f$ belong to H^2 and ∇f belongs to L^2_2.

Proposition 2.2 relies on the next lemma, whose proof is elementary and hence omitted.

LEMMA 2.3. *Let $(m,k) \in \mathbb{N}^2$, with $m \leq k$. If $f \in H^m_k$, then $(1+|x|^k)\,f \in H^m$.*

Thanks to Lemma 2.3, we can prove Proposition 2.2.

PROOF OF PROPOSITION 2.2. Assertion (i) is trivial. So, let us focus on the proof of (ii).

Suppose that $f \in H^m \cap H^{m-1}_k$. Then, by Lemma 2.3 above, we have that $|x|^k\,f \in H^{m-1}_k$. At this point, for $|\alpha| = m$, we write, using Leibniz rule,

$$D^\alpha\Big(|x|^k\,f\Big) \,=\, |x|^k\,D^\alpha f \,+\, \sum_\beta D^\beta |x|^k\,D^{\alpha-\beta} f\,,$$

where the sum is performed for all $\beta \leq \alpha$ such that $|\beta| \geq 1$. By the previous arguments, and the fact that $m \leq k$, we have that all the terms in the sum belong to L^2. Then, the term on the left-hand side belongs to L^2 if and only if the first term on the right-hand side does.

The last sentences follow by straightforward computations, using the equality $\partial_j\big(|x|\,f\big) \,=\, \partial_j|x|\,f \,+\, |x|\,\partial_j f$, where $1 \leq j \leq d$, and the relation

$$\nabla^2\Big(|x|^2\,f\Big) \sim \nabla\Big(|x|\,f + |x|^2\,\nabla f\Big) \sim \nabla|x|\,f + \Big(|x| + |x|^2\Big)\,\nabla f + |x|^2\,\nabla^2 f\,.$$

The equivalence between the two assertions is then apparent. Indeed, we have that if $f \in H^2_2$, then $|x|^j\,D^\alpha f \in L^2$ for all $0 \leq j \leq 2$ and $|\alpha| = 0, 1$. Hence, all the terms in the right-hand side belong to L^2, and then so does the one on the left-hand side.

On the contrary, if both f and $|x|^2\,f$ belong to H^2 and ∇f belongs to L^2_2, then $f \in H^2 \cap H^1_2$; finally, by the previous equality, we also discover that $|x|^2\,\nabla^2 f$ belongs to L^2, completing the proof of the reverse implication, and then of the whole proposition. □

After the above preliminaries, we are ready to state the main result of this section, which shows well-posedness of the Liouville equation in H^m_k spaces.

Theorem 2.4. *Let $T>0$ and $(m,k)\in\mathbb{N}^2$ fixed, and let a be a vector field satisfying hypotheses* (2.11). *Assume also that $f_0\in H^m_k(\mathbb{R}^d)$ and $g\in L^1\Big([0,T];H^m_k(\mathbb{R}^d)\Big)$.*

Then there exists a unique solution $f\in C\Big([0,T];H^m_k(\mathbb{R}^d)\Big)$ to problem (2.9). *Moreover, there exists a "universal" constant $C>0$, independent of f_0, a, g, f and T, such that the following estimate holds true for any $t\in[0,T]$:*

$$\|f(t)\|_{H^m_k}\,\leq\,C\,\exp\left(C\int_0^t\|\nabla a(\tau)\|_{C^m_b}\,d\tau\right)\,\left(\|f_0\|_{H^m_k}\,+\,\int_0^t\|g(\tau)\|_{H^m_k}\,d\tau\right)\,. \tag{2.24}$$

Most of the claims of the previous statement follow from Theorem 2.2. What remains is to prove the propagation of higher integrability for $k\geq1$. Before proving Theorem 2.4 in its full generality, let us consider its version for simpler cases, which are needed in the proof of the general case. Moreover, their precise form is important, in view of their application in Section 3.2.

We start with the case $m=0$.

Lemma 2.4. *Assume that the hypotheses of Theorem 2.4 hold true with $m=0$.*

Then there exists a unique solution $f\in C\Big([0,T];L^2_k(\mathbb{R}^d)\Big)$ to problem (2.9). *Moreover, there exists a "universal" constant $C>0$ such that the following estimate holds true for any $t\in[0,T]$*

$$\|f(t)\|_{L^2_k}\,\leq\,C\,\exp\left(C\int_0^t\|\nabla a(\tau)\|_{L^\infty}\,d\tau\right)\,\left(\|f_0\|_{L^2_k}\,+\,\int_0^t\|g(\tau)\|_{L^2_k}\,d\tau\right)\,.$$

Proof of Lemma 2.4. Recall that, in the case $k=0$, taking the L^2 scalar product of equation (2.9) by f leads to

$$\frac{1}{2}\,\frac{\mathrm{d}}{\mathrm{d}t}\,\|f\|^2_{L^2}\,+\,\frac{1}{2}\int\mathrm{div}\,a\,|f|^2\,dx\,=\,\int g\,f\,dx\,,$$

which readily implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\|f\|_{L^2}\,\leq\,\|\,\mathrm{div}\,a\|_{L^\infty}\,\|f\|_{L^2}\,+\,\|g\|_{L^2}\,. \tag{2.25}$$

Analogously, multiplying equation (2.9) by $|x|^k$, we get that $f_k:=|x|^k\,f$ satisfies

$$\partial_tf_k\,+\,\mathrm{div}\,\Big(a\,f_k\Big)\,=\,|x|^k\,g\,+\,f\,a\cdot\nabla|x|^k\,.$$

Taking now the L^2 scalar product by f_k and repeating the same computations as above, we find

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\|f_k\|_{L^2}\,\leq\,\|\,\mathrm{div}\,a\|_{L^\infty}\,\|f_k\|_{L^2}\,+\,\Big\||x|^k\,g\Big\|_{L^2}\,+\,\Big\|f\,a\cdot\nabla|x|^k\Big\|_{L^2}\,. \tag{2.26}$$

We need to control the last term on the right-hand side of the previous estimate. To achieve this, we use the fact that $\nabla|x|^k\,\sim\,|x|^{k-1}$ for all $k\geq1$ and Remark 2.2 to obtain

$$\left\| f\, a \cdot \nabla |x|^k \right\|_{L^2} \,\leq\, C\, \|\nabla a\|_{L^\infty} \, \left\| \left(1 + |x|^k\right) f \right\|_{L^2} .$$

Inserting this bound into (2.26) and summing up the resulting expression to (2.25), we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \left\| \left(1 + |x|^k\right) f \right\|_{L^2} \,\leq\, C\, \|\nabla a\|_{L^\infty} \, \left\| \left(1 + |x|^k\right) f \right\|_{L^2} \,+\, \left\| \left(1 + |x|^k\right) g \right\|_{L^2} . \tag{2.27}$$

Hence, an application of Grönwall's lemma gives the desired estimate. □

Next, we present the result for $m = 1$. For notational convenience, let us set

$$[x]_k \,:=\, 1 \,+\, |x|^k .$$

Lemma 2.5. *Assume that the hypotheses of Theorem 2.4 hold true with $m = 1$. Then there exists a unique weak solution $f \in C\Big([0,T]; H^1_k(\mathbb{R}^d)\Big)$ to (2.9). Moreover, there exists a "universal" constant $C > 0$ such that the estimate*

$$\|f(t)\|_{H^1_k} \,\leq\, C \,\exp\left(C \int_0^t \|\nabla a(\tau)\|_{C^1_b}\, d\tau\right) \left(\|f_0\|_{H^1_k} \,+\, \int_0^t \|g(\tau)\|_{H^1_k}\, d\tau\right) ,$$

is satisfied for all $t \in [0,T]$.

Proof of Lemma 2.5. We start by differentiating equation (2.9) with respect to x^j, for some $1 \leq j \leq d$, getting

$$\partial_t \partial_j f \,+\, \mathrm{div}\left(a\, \partial_j f\right) \,=\, \partial_j g \,-\, \partial_j \,\mathrm{div}\, a\; f \,-\, \partial_j a \cdot \nabla f .$$

Applying estimate (2.27) to this equation gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \|[x]_k\, \partial_j f\|_{L^2} \,\leq\, C\, \|\nabla a\|_{L^\infty}\, \|[x]_k\, \partial_j f\|_{L^2} \,+\, \|[x]_k\, \partial_j g\|_{L^2} \\ +\, \|[x]_k\, \partial_j \,\mathrm{div}\, a\; f\|_{L^2} \,+\, \|[x]_k\, \partial_j a \cdot \nabla f\|_{L^2} \,,\end{aligned}$$

from which we obtain, for another constant $C > 0$, the bound

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \|[x]_k\, \nabla f\|_{L^2} \,\leq\, C\, \|\nabla a\|_{L^\infty}\, \|[x]_k\, \nabla f\|_{L^2} \,+\, \|[x]_k\, \nabla g\|_{L^2} \\ +\, \left\|\nabla^2 a\right\|_{L^\infty}\, \|[x]_k\, f\|_{L^2} .\end{aligned} \tag{2.28}$$

We can now sum up (2.27) and (2.28) to get

$$\frac{\mathrm{d}}{\mathrm{d}t} \|f\|_{H^1_k} \,\leq\, C\, \|\nabla a\|_{C^1_b}\, \|f\|_{H^1_k} \,+\, \|g\|_{H^1_k} . \tag{2.29}$$

An application of Grönwall's lemma allows us to get the result. □

Now, we can address the proof of the general case given in Theorem 2.4.

Proof of Theorem 2.4. We argue by induction on the order of derivatives. The cases $m = 0$ and $m = 1$ are given by Lemma 2.4 and Lemma 2.5, respectively.

Let $m \geq 2$, and let us assume that, for any $0 \leq \ell \leq m-1$, the following inequality holds true

$$\frac{\mathrm{d}}{\mathrm{d}t} \left\| [x]_k \, \nabla^\ell f \right\|_{L^2} \leq C \|\nabla a\|_{L^\infty} \left\| [x]_k \, \nabla^\ell f \right\|_{L^2} + \left\| [x]_k \, \nabla^\ell g \right\|_{L^2} + \sum_{0 \leq p \leq \ell - 1} \left\| \nabla^{p+1} a \right\|_{L^\infty} \| [x]_k \, \nabla^p f \|_{L^2} \, . \tag{2.30}$$

Our goal is to prove an analogous estimate also for $\| [x]_k \, \nabla^m f \|_{L^2}$.

For this purpose, let us take an $\alpha \in \mathbb{N}^d$ such that $|\alpha| = m$. Applying the operator D^α to (2.9), we deduce

$$\partial_t D^\alpha f + \mathrm{div}\left(a \, D^\alpha f \right) = D^\alpha g - \sum_{0 < \beta \leq \alpha} D^\beta \, \mathrm{div}\, a \; D^{\alpha - \beta} f - \sum_{0 < \beta \leq \alpha} D^\beta a \cdot \nabla D^{\alpha - \beta} f \, , \tag{2.31}$$

where the notation $0 < \beta$ means that $\beta \in \mathbb{N}^d$ has at least one non-zero component.

Following the computations of Lemma 2.5, we need to estimate the L^2_k norm of the last two terms in the right-hand side of the previous equation. For the first one, we have

$$\left\| [x]_k \, D^\beta \, \mathrm{div}\, a \; D^{\alpha - \beta} f \right\|_{L^2} \leq \left\| \nabla^{|\beta| + 1} a \right\|_{L^\infty} \left\| [x]_k \, D^{\alpha - \beta} f \right\|_{L^2} \, .$$

Notice that, since $\beta > 0$, the terms $D^{\alpha - \beta} f$ are of lower order. It holds for the terms

$$\left\| [x]_k \, D^\beta a \cdot \nabla D^{\alpha - \beta} f \right\|_{L^2} \leq \left\| \nabla^{|\beta|} a \right\|_{L^\infty} \left\| [x]_k \, \nabla D^{\alpha - \beta} f \right\|_{L^2} \, ,$$

whenever $|\beta| \geq 2$; on the contrary, when $|\beta| = 1$, the terms $\nabla D^{\alpha - \beta} f$ contain exactly m derivatives.

Therefore, applying estimate (2.27) to equation (2.31), and using the previous controls, we infer

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \| [x]_k \, \nabla^m f \|_{L^2} &\leq C \|\nabla a\|_{L^\infty} \; \| [x]_k \, \nabla^m f \|_{L^2} + \| [x]_k \, \nabla^m g \|_{L^2} \\ &\qquad + \sum_{0 < \beta \leq \alpha} \left\| \nabla^{|\beta| + 1} a \right\|_{L^\infty} \left\| [x]_k \, D^{\alpha - \beta} f \right\|_{L^2} \\ &\leq C \|\nabla a\|_{L^\infty} \; \| [x]_k \, \nabla^m f \|_{L^2} + \| [x]_k \, \nabla^m g \|_{L^2} \\ &\qquad + \sum_{0 \leq \ell \leq m-1} \left\| \nabla^{\ell + 1} a \right\|_{L^\infty} \left\| [x]_k \, \nabla^\ell f \right\|_{L^2} \, , \end{aligned}$$

which proves formula (2.30) at the level m. Now, it is a matter of summing up inequality (2.28) for $\ell = 0$ to m to get, for some constant also depending on m, the following bound:

$$\frac{\mathrm{d}}{\mathrm{d}t} \|f\|_{H^m_k} \leq C \; \|\nabla a\|_{C^m_b} \; \|f\|_{H^m_k} + \|g\|_{H^m_k} \, ,$$

which immediately implies the claimed estimate. Theorem 2.4 is now proved. □

2.3. Numerical analysis of the Liouville equation

In this section, we present our strategy and analysis of the discretization of the controlled Liouville equation (2.6) and its optimization adjoint model (2.8). In the adjoint model, we fix the initial condition $q(x,0) = -\varphi(x)$ and the source term $g(x,t) = -\theta(x,t)$ for $(x,t) \in \Omega \times [0,T]$, where θ and φ are Gaussian functions; cf. (3.32).

We begin with discussing the spatial and temporal discretization of the Liouville equation and its adjoint. For simplicity of notation, we focus on a two-dimensional problem, i.e. $d = 2$. Then $a = (a^1, a^2) \in \mathbb{R}^2$. For the numerical analysis of the Liouville equation, we fix the drift function as in (2.5). Moreover, we fix the controls to belong to

$$U_{ad} := \left\{ u \in \mathbb{L}^\infty_T(\mathbb{R}^d) \;\middle|\; u^a \leq u(t) \leq u^b \qquad \text{for a.e. } t \in [0,T] \right\}, \tag{2.32}$$

where the inequalities are meant component-wise, and we choose $u^a = \left(u_1^a, u_2^a\right)$ and $u^b = \left(u_1^b, u_2^b\right)$ in $\mathbb{R}^{2d}$, with $u^a < u^b$.

Our aim is to develop an approximation framework that is second-order accurate and preserves the two essential properties of the continuous Liouville model given in (2.4), namely positivity and conservation of mass of its solution.

In view of applications to the numerical study of our optimal control problem, we consider a large but bounded convex domain $\Omega \subset \mathbb{R}^2$. Specifically, we choose $\Omega = (-B,B) \times (-B,B)$, for some large $B > 0$. We also fix an initial density f_0 that is (by machine precision) compactly supported in Ω. For θ and φ we take Gaussian functions having sufficiently small variance and centred sufficiently far from the boundary of Ω, so that (by machine precision) we can assume that also those functions are compactly supported in Ω. Then, we solve problems (2.6) and (2.8) in $\Omega \times [0,T]$, supplemented with homogeneous Dirichlet boundary conditions on $\partial\Omega$. Notice that, in this setting, it is possible to use the results of the current chapter to prove existence and uniqueness of smooth enough solutions to (2.6) and (2.8). For this purpose, one extends the functions f_0, θ and φ to be zero outside the domain Ω, and the drift function a to a smooth function, which is bounded on $\mathbb{R}^d$ together with all its space derivatives. We remark that this is always possible, for instance by multiplying a with a smooth compactly supported function χ of the space variable only, such that $\chi \equiv 1$ on a neighbourhood of Ω.

We consider our solutions on a time interval $[0,T]$, where $T > 0$ is chosen such that the corresponding solution f to (2.6) is still compactly supported in Ω, far away from its boundary $\partial\Omega$. Observe that this property is true by finite propagation speed, since the (extended) drift is bounded on $\mathbb{R}^d$. Also notice that, in our analysis, we do not need the solution q to the adjoint problem to be compactly supported in Ω.

For $\Omega = (-B, B) \times (-B, B)$ fixed above, we set a numerical grid that provides a partitioning of Ω in $N_x \times N_x$, $N_x > 1$, equally-spaced non-overlapping square cells of side length $h = 2B/N_x$. On this partitioning, we consider a cell-centred finite-volume setting, where the nodal points at which the density and adjoint variables are defined are placed at the centres of the square volumes. These nodal points are given by

$$x_1^i := \left(i - \frac{1}{2}\right) h - B, \qquad x_2^j := \left(j - \frac{1}{2}\right) h - B. \tag{2.33}$$

Therefore, the elementary cell is defined as

$$\omega_h^{ij} := \left\{ (x_1, x_2) \in \Omega \;\middle|\; x_1 \in \left[x_1^i - \frac{h}{2}, x_1^i + \frac{h}{2}\right], \quad x_2 \in \left[x_2^j - \frac{h}{2}, x_2^j + \frac{h}{2}\right] \right\}, \tag{2.34}$$

and the computational domain is given by

$$\Omega_h = \bigcup_{i,j=1}^{N_x} \omega_h^{ij}. \tag{2.35}$$

Analogously, the time interval $[0, T]$ is divided in $N_t > 1$ subintervals of length Δt and the points t^k are given by

$$t^k := k\Delta t, \qquad k = 0, \ldots, N_t, \qquad \Delta t := \frac{T}{N_t}.$$

This defines the time mesh $\Gamma_{\Delta t} := \{t^k \in [0, T],\ k = 0, \ldots, N_t\}$. Therefore, corresponding to the space-time cylinder $\mathcal{Z} := \Omega \times [0, T]$ we have its discrete counterpart $\mathcal{Z}_{h,\Delta t} := \Omega_h \times \Gamma_{\Delta t}$.

In this setting, the cell average of the density f (and so of any integrable function), on the cell with centre (x_1^i, x_2^j) at time t^k, is given by

$$\bar{f}_{i,j}^k = \frac{1}{h^2} \int_{x_1^{i-1/2}}^{x_1^{i+1/2}} \int_{x_2^{j-1/2}}^{x_2^{j+1/2}} f(x_1, x_2, t^k)\, dx_2\, dx_1. \tag{2.36}$$

In particular,

$$\bar{f}_{i,j}^0 = f_{i,j}^0 = \frac{1}{h^2} \int_{x_1^{i-1/2}}^{x_1^{i+1/2}} \int_{x_2^{j-1/2}}^{x_2^{j+1/2}} f_0(x_1, x_2)\, dx_2\, dx_1.$$

Notice that, in our numerical setting, function values are identified with their cell-based average located at the cell centres. For this reason, our numerical framework aims at determining approximation of theses averages. Specifically, we discuss a discretization scheme that results in values $f_{i,j}^k$ that approximate $\bar{f}_{i,j}^k$. Similarly, we denote with $q_{i,j}^k$ the numerical approximation of $\bar{q}_{i,j}^k$ that is computed as in (2.36).

We also consider a piecewise constant approximation to the time-dependent control functions, where we denote with $u^{k+1/2}$ the value of the control in the time interval

$[t^k, t^{k+1})$. Further, for the projection of a continuous u to the corresponding approximation space, we set $u^{k+1/2} = u(t^k)$. For a function f defined on $\mathcal{Z}_{h,\Delta t}$, we define the discrete norms $\|\cdot\|_{1,h}$ and $\|\cdot\|_{\infty,h}$ as follows:

$$\|f(\cdot,\cdot,t^k)\|_{1,h} = h^2 \sum_{i,j}^{N_x} \left|f_{i,j}^k\right|, \qquad \|f(\cdot,\cdot,t^k)\|_{\infty,h} = \max_{i,j=1,\ldots,N_x} \left|f_{i,j}^k\right|,$$

where $f_{i,j}^k = f(x_1^i, x_2^j, t^k)$, and (x_1^i, x_2^j, t^k) denotes a grid point in $\Omega \times [0,T]$.

2.3.1. A Runge-Kutta Kurganov-Tadmor scheme for the Liouville equation

In this section, we discuss a suitable approximation of our controlled Liouville equation in $\Omega \times [0,T]$. Supposing that f_0 has compact support, and because of finite propagation speed, we can choose Ω such that the solution f at the boundary $\partial\Omega$ is zero for all times $t \in [0,T]$.

For our purpose, we focus on the finite-volume scheme proposed by Kurganov-Tadmor (KT) in [**92**] that involves a generalized MUSCL flux. To describe this scheme, we denote the flux in the Liouville equation as a variable of f with $\mathcal{H}(f) = a\,f = a(x,t)\,f(x,t)$. Thus, the KT scheme for the Liouville equation in semi-discretized form is given by

$$\text{(2.37)} \quad \frac{\mathrm{d}}{\mathrm{d}t} f_{i,j}(t) = -\frac{H_{i+1/2,j}^{x_1}(f^+,f^-;t) - H_{i-1/2,j}^{x_1}(f^+,f^-;t)}{h} - \frac{H_{i,j+1/2}^{x_2}(f^+,f^-;t) - H_{i,j-1/2}^{x_2}(f^+,f^-;t)}{h}, \quad i,j = 1,\ldots,N_x - 1,$$

where the $H_{\cdot,\cdot}^{x_r}(f^+,f^-;t)$ are the fluxes in r-direction, $r = 1,2$. Specifically, for $H_{\cdot,\cdot}^{x_1}(f^+,f^-;t)$ we have

$$\text{(2.38)} \quad H_{i+1/2,j}^{x_1}(f^+,f^-;t) := \frac{\mathcal{H}^1(f_{i+1/2,j}^+(t)) + \mathcal{H}^1(f_{i+1/2,j}^-(t))}{2} - \frac{\mathcal{V}_{i+1/2,j}^{x_1}(t)}{2}\Big[f_{i+1/2,j}^+(t) - f_{i+1/2,j}^-(t))\Big],$$

where $\mathcal{H} = (\mathcal{H}^1, \mathcal{H}^2) = (a^1 f, a^2 f)$, and similarly for $H_{i,j\pm1/2}^{x_2}(f^+,f^-;t)$. In this formula, the so-called local speeds $\mathcal{V}^{x_r}(t)$ are given by

$$\text{(2.39)} \qquad \mathcal{V}_{i+1/2,j}^{x_r}(t) = \left| a^r(x_1^{i+1/2}, x_2^j, t; u(t)) \right|, \qquad r = 1,2,$$

since $\mathcal{H}(f) = a\,f$ is linear in f.

Further in (2.38), the approximation of f at the cell edges is given by intermediate values that approximate the function value from above respectively from below as follows

$$\text{(2.40)} \quad f_{i+1/2,j}^+(t) := f_{i+1,j}(t) - \frac{h}{2}(f_{x_1})_{i+1,j}(t), \qquad f_{i+1/2,j}^-(t) := f_{i,j}(t) + \frac{h}{2}(f_{x_1})_{i,j}(t).$$

We approximate the partial derivatives of f using the minmod function. In direction x_1, this approximation is given by

(2.41)
$$(f_{x_1})_{i,j}(t) = \operatorname{minmod}\Big(\frac{f_{i,j}(t) - f_{i-1,j}(t)}{h}, \frac{f_{i+1,j}(t) - f_{i-1,j}(t)}{2h}, \frac{f_{i+1,j}(t) - f_{i,j}(t)}{h}\Big).$$

An analogous expression holds in the direction x_2. The multivariable minmod function for vectors $x \in \mathbb{R}^d$ is given by

$$\operatorname{minmod}(x_1, x_2, \dots, x_d) := \begin{cases} \min_j\{x_j\} & \text{if } x_j > 0,\ \forall j \in [1,d] \\ \max_j\{x_j\} & \text{if } x_j < 0,\ \forall j \in [1,d] \\ 0 & \text{otherwise.} \end{cases}$$

Next, we discuss the local truncation error of the semi-discrete KT scheme (2.37)–(2.39).

Lemma 2.6. *The KT scheme given in (2.37)–(2.40) is at least second-order accurate for smooth f, except possibly at the points of extrema of f.*

Proof. The flux H, given in (2.38), is a first-order Rusanov flux [**113**] that is C^2 with Lipschitz continuous partial derivatives with respect to f^+, f^-, in a neighbourhood of $f_{i,j}$. Further, by using a Taylor series expansion, we have the following approximation

$$h\frac{(f_{x_1})_{i+1,j}}{f_{i+1,j} - f_{i,j}} = 1 + \mathcal{O}(h), \qquad h\frac{(f_{x_1})_{i,j}}{f_{i+1,j} - f_{i,j}} = 1 + \mathcal{O}(h),$$

and a similar result holds in the x_2 direction, except at the points of extrema, which are characterized by $(f_{x_1})_{i,j} = 0$ (see [**104**, Th. 3.2]). Using the result in [**103**, Lemma 2.1], we have that the semi-discrete scheme (2.37)–(2.40) is second-order accurate in space except possibly at points of extrema. □

For the time discretization of the Liouville equation (3.6), we use a second-order strong stability preserving Runge-Kutta (SSPRK2) method [**114**] (also known as the Huen's method). The combination of this scheme with the KT discretization of the flux $\mathcal{H}$, given in (2.37), results in a new approximation method that we call the SSPRK2-KT-scheme. This scheme is implemented by the Algorithm 2.1 given below using the following definition

(2.42)
$$F(f^k_{i,j}) = -\frac{H^{x_1,k}_{i+1/2,j} - H^{x_1,k}_{i-1/2,j}}{h} - \frac{H^{x_2,k}_{i,j+1/2} - H^{x_2,k}_{i,j-1/2}}{h}.$$

where $H^{x_j,k}_{\cdot,\cdot}$ denotes $H^{x_j}_{\cdot,\cdot}$, $j = 1, 2$, given in (2.38) corresponding to the time-step t^k.

Algorithm 2.1 SSPRK2-KT scheme

```
Require: $f^0_{\cdot,\cdot}$, $F$
Ensure: Solve the Liouville equation in $f$ as follows
Set $k = 0$
while $0 \le k < N_t$ do
    for $1 < i, j < N_x - 1$ do
        In $(t^k, t^{k+1})$, compute $f^{(1)}_{i,j} = f^k_{i,j} + \Delta t\, F(f^k_{i,j})$ with initial condition $f^k_{i,j}$, where $F(f^k_{i,j})$ is computed using (2.42).
        In $(t^k, t^{k+1})$, compute $f^{(2)}_{i,j} = f^{(1)}_{i,j} + \Delta t\, F(f^{(1)}_{i,j})$ with initial condition $f^{(1)}_{i,j}$, where $F(f^{(1)}_{i,j})$ is computed using (2.42).
        Time-step update: $f^{k+1}_{i,j} = \frac{1}{2} f^k_{i,j} + \frac{1}{2} f^{(2)}_{i,j}$.
    end for
    $k = k + 1$
end while
return $f^k_{\cdot,\cdot}$
```

Now, we study the properties of the SSPRK2-KT scheme given in Algorithm 2.1. This method has the following strong stability property [**77**, Lemma 2.1]

$$\|f^{k+1}\|_{\infty,h} \le \|f^k\|_{\infty,h}, \qquad k = 0, \dots, N_t - 1.$$

Further, we have conservation of the total probability (or mass) as a consequence of the finite-volume formulation:

Lemma 2.7 (Conservation). *The SSPRK2-KT scheme is conservative, in the sense that*

$$\sum_{i,j=1}^{N_x} f^k_{i,j} = \sum_{i,j=1}^{N_x} f^0_{i,j}, \qquad k = 1, \dots, N_t.$$

Proof. Fix $k \in \{0, \dots, N_t\}$. From the first step of the SSPRK2-KT scheme, given in Algorithm 2.1, summing up over all indices $i, j \in \{1, \dots, N_x\}$ and using the zero flux boundary condition (**??**), we have

$$\sum_{i,j=1}^{N_x} f^{(1)}_{i,j} = \sum_{i,j=1}^{N_x} f^k_{i,j}.$$

In a similar way, we have

$$\sum_{i,j=1}^{N_x} f^{(2)}_{i,j} = \sum_{i,j=1}^{N_x} f^{(1)}_{i,j} = \sum_{i,j=1}^{N_x} f^k_{i,j}.$$

Thus,

$$\sum_{i,j=1}^{N_x} f^{k+1}_{i,j} = \frac{1}{2} \sum_{i,j=1}^{N_x} f^k_{i,j} + \frac{1}{2} \sum_{i,j=1}^{N_x} f^{(2)}_{i,j} = \sum_{i,j=1}^{N_x} f^k_{i,j}.$$

Iterating over k, we have

$$\sum_{i,j=1}^{N_x} f_{i,j}^k = \sum_{i,j=1}^{N_x} f_{i,j}^0, \qquad k = 1, \ldots, N_t.$$

□

Next, we show that, starting from a non-negative initial density, the solution obtained with the SSPRK2-KT scheme remains non-negative. For this purpose, we define the CFL-number

$$\lambda := \frac{\Delta t}{h}, \tag{2.43}$$

and require the conditions on the components of the drift a given in (2.5) given by

$$\lambda \left\| a^1 \right\|_{L_T^\infty(L^\infty(\Omega))} \le \frac{1}{4}, \qquad \lambda \left\| a^2 \right\|_{L_T^\infty(L^\infty(\Omega))} \le \frac{1}{4}.$$

Notice that the control u belongs to the set U_{ad} defined in (3.23). Then, for $j = 1, 2$, we see that

$$\left\| a^j \right\|_{L_T^\infty(L^\infty(\Omega))} \le \left\| a_0^j \right\|_{L_T^\infty(L^\infty(\Omega))} + \Big(b + c\,B\Big) \max\Big\{|u^a|, |u^b|\Big\} \ ,$$

so that the aforementioned conditions on the components of the drift a are satisfied under the following CFL condition.

$$\lambda \left(\left\| a_0^j \right\|_{L_T^\infty(L^\infty(\Omega))} + \Big(b + c\,B\Big) \max\Big\{|u^a|, |u^b|\Big\} \right) \le \frac{1}{4}\,, \qquad j = 1, 2\,. \tag{2.44}$$

This CFL condition only depends on the components of the vector a_0, and it does not rest on the unknowns of the problem.

With the conditions (2.44), we can prove the following lemma on the positivity of the SSPRK2-KT scheme.

Lemma 2.8 (Positivity). *Under the CFL-condition (2.44), with $u \in U_{ad}$ and $f_{i,j}^0 \ge 0$, $i, j = 1, \ldots, N_x$, the solution to the Liouville problem computed with the SSPRK2-KT scheme given in Algorithm 2.1 is non-negative, that is,*

$$f_{i,j}^k \ge 0, \qquad i, j = 1, \ldots, N_x, \qquad k = 1, \ldots, N_t. \tag{2.45}$$

Proof. The SSPRK2-KT scheme, given in Algorithm 2.1, comprises of a two-step Euler scheme that results in the computation of $f^{(1)}$ and $f^{(2)}$ and a final averaging step. To prove positivity of the SSPRK2-KT scheme, it is enough to show that the solutions obtained in each of the two Euler steps is positive. Without loss of generality, we prove that the solution obtained in the first step of the SSPRK2-KT scheme is positive. A similar analysis holds true for the second step.

Let $f_{i,j}^k \geq 0$ for fixed $0 \leq k < N_t$. We will show that $f_{i,j}^{k+1} \geq 0$ for all $i, j = 1, \ldots, N_x$. For this purpose, notice that the SSPRK2-KT scheme can be written as follows

(2.46)

$$\begin{aligned} f_{i,j}^{k+1} = &\frac{\lambda}{2}\Big(|a^1_{i+1/2,j}| - a^1_{i+1/2,j}\Big) f^+_{i+1/2,j} + \frac{\lambda}{2}\Big(|a^1_{i-1/2,j}| + a^1_{i-1/2,j}\Big) f^-_{i-1/2,j} \\ &+\frac{\lambda}{2}\Big(|a^2_{i,j+1/2}| - a^2_{i,j+1/2}\Big) f^+_{i,j+1/2} + \frac{\lambda}{2}\Big(|a^2_{i,j-1/2}| + a^2_{i,j-1/2}\Big) f^-_{i,j-1/2} \\ &+\left[\frac{1}{4} - \frac{\lambda}{2}\Big(|a^1_{i+1/2,j}| + a^1_{i+1/2,j}\Big)\right] f^-_{i+1/2,j} + \left[\frac{1}{4} - \frac{\lambda}{2}\Big(|a^1_{i-1/2,j}| - a^1_{i-1/2,j}\Big)\right] f^+_{i-1/2,j} \\ &+\left[\frac{1}{4} - \frac{\lambda}{2}\Big(|a^2_{i,j+1/2}| + a^2_{i,j+1/2}\Big)\right] f^-_{i,j+1/2} + \left[\frac{1}{4} - \frac{\lambda}{2}\Big(|a^2_{i,j-1/2}| - a^2_{i,j-1/2}\Big)\right] f^+_{i,j-1/2}, \end{aligned}$$

where all discrete quantities on the right-hand side are considered at the time-step t^k. We see that the first four terms on the right hand side in (2.46) are always non-negative, provided that $f^{\pm}_{i\pm1/2,j}$, $f^{\pm}_{i,j\pm1/2} \geq 0$. The remaining terms are non-negative under the CFL-condition (2.44).

Thus, it remains to show that $f^{\pm}_{i+1/2,j}$, $f^{\pm}_{i,j+1/2} \geq 0$ for all $i, j = 1, \ldots, N_x$, where $f^{\pm}_{i,j}$ is given as in (2.40).

We consider each of the expressions for $(f_{x_1})^k_{i,j}$ in the direction of x_1 given as in (2.41). First, assume that $(f_{x_1})^k_{i,j} = \frac{f^k_{i,j} - f^k_{i-1,j}}{h}$, which is one of the possible values of the minmod limiter in (2.41).

Then it follows that

$$f^+_{i+1/2,j} = \left(1 - \frac{1}{2}\right) f^k_{i+1,j} + \frac{1}{2} f^k_{i,j}.$$

This is non-negative, since $f^k_{i,j} \geq 0$ for all $i, j = 1, \ldots, N_x$ by the inductive assumption. Further, $f^-_{i+1/2,j} = f^k_{i,j} + \frac{h}{2}\left[\frac{f^k_{i,j} - f^k_{i-1,j}}{h}\right]$. If $\frac{f^k_{i,j} - f^k_{i-1,j}}{h} > 0$, then it implies $f^-_{i+1/2,j} > 0$. If $\frac{f^k_{i,j} - f^k_{i-1,j}}{h} < 0$, then by the definition of the minmod limiter, we have $\frac{f^k_{i,j} - f^k_{i-1,j}}{h} \geq \frac{f^k_{i+1,j} - f^k_{i,j}}{h}$ and therefore

$$f^-_{i+1/2,j} \geq f^k_{i,j} + \frac{h}{2}\left[\frac{f^k_{i+1,j} - f^k_{i,j}}{h}\right] = \frac{f^k_{i+1,j} + f^k_{i,j}}{2} \geq 0.$$

The other cases for the value of $(f_x)^k_{i,j} \neq 0$ follow analogously. If $(f_x)^k_{i,j} = 0$, then $f^{\pm}_{i+1/2,j} = f_{i+1,j} \geq 0$ and $f^{\pm}_{i,j+1/2} = f_{i,j+1} \geq 0$. This proves the lemma. □

Remark 2.3. *The proof of the above lemma follows similar arguments as in* [**91**, Theorem 2.1]. *However, a primary difference is that in* [**91**], *the positivity result is proved using one-sided local speeds, exploiting the structure of the hyperbolic equation. In our case, the proof relies on conversion of the intermediate values $f^{\pm}$ to the cell-average values $f^k_{i,j}$ and then showing that $f^k_{i,j} \geq 0$ implies $f^{k+1}_{i,j} \geq 0$, which seems a much simpler approach.*

Remark 2.4. *Under the same CFL-like condition* (2.44)*, the proof of Lemma 2.8 can be extended to the case of a SSPRK-KT scheme with a Runge-Kutta method of p-th order, $p \in \mathbb{N}$, that is given as an average of p Euler steps.*

Remark 2.5. *For the case where $(f_{x_1})^k_{i,j} = 0$, the approximations of f at the cell-edges, given by* (2.40)*, are piecewise constant in the cell ω_h^{ij}. Thus, the KT scheme given by* (2.37)–(2.40)*, reduces to a linear upwind scheme that is locally first-order accurate, TVD and positive. This is consistent with the Godunov's barrier theorem.*

Next, we prove discrete L^1 stability of the SSPRK2-KT scheme.

Lemma 2.9 (Stability). *The solution $f^k_{i,j}$ obtained with the SSPRK2-KT-scheme in Algorithm 2.1 is discrete L^1 stable in the sense that*

$$\left\|f^k_{\cdot,\cdot}\right\|_{1,h} = \left\|f^0_{\cdot,\cdot}\right\|_{1,h}, \qquad k = 1, \dots, N_t,$$

under the CFL condition (2.44).

Proof. Using Lemma 2.7, we have

$$\sum_{i,j=0}^{N_x} f^k_{i,j} = \sum_{i,j=0}^{N_x} f^0_{i,j}, \qquad k = 1, \dots, N_t.$$

Again, from Lemma 2.8, we have that the solution obtained with the SSPRK2-KT scheme is positive under the CFL condition (2.44). This gives us the following relation

$$\sum_{i,j=0}^{N_x} |f^k_{i,j}| = \sum_{i,j=0}^{N_x} |f^0_{i,j}|, \qquad k = 1, \dots, N_t,$$

which proves the result. □

Next, we aim at proving the L^1 convergence of the SSPRK2-KT scheme. For this purpose, we prove the following stability estimate for the discrete solution of the Liouville equation (3.6) with a right-hand side function $g(x,t)$.

Lemma 2.10. *Let $f^k_{i,j}$ be the SSPRK2-KT solution to the Liouville equation* (3.6) *with a Lipschitz continuous right-hand side $g(x,t)$ and let the CFL condition* (2.44) *be fulfilled. Then this solution satisfies the following stability estimate*

$$\left\|f^{k+1}_{\cdot,\cdot}\right\|_{1,h} \leq \left\|f^0_{\cdot,\cdot}\right\|_{1,h} + \Delta t \sum_{m=0}^{k} \left\|g^m_{\cdot,\cdot}\right\|_{1,h},$$

for $k = 1, \dots, N_t - 1$ where $g^m_{i,j} = g(x_1^i, x_2^j, t^m)$.

Proof. The SSPRK2-KT scheme, given in Algorithm 2.1, for the Liouville equation (3.6) with a right-hand side $g(x,t)$ can be written as

$$\begin{aligned}\frac{f^{(1)}_{i,j}-f^k_{i,j}}{\Delta t}&=-\frac{1}{2h}(L^k_i+L^k_j)(f)+g^k_{i,j},\\ \frac{f^{k+1}_{i,j}-f^k_{i,j}}{\Delta t}&=-\frac{1}{4h}(L^k_i+L^k_j+L^{(1)}_i+L^{(1)}_j)(f)+g^k_{i,j},\end{aligned}\tag{2.47}$$

where

$$\begin{aligned}L^n_i(f)=&\Big(|a^1_{i+1/2,j}|-a^1_{i+1/2,j}\Big)f^{n+}_{i+1/2,j}-\Big(|a^1_{i+1/2,j}|+a^1_{i+1/2,j}\Big)f^{n-}_{i+1/2,j}\\ &+\Big(|a^1_{i-1/2,j}|+a^1_{i-1/2,j}\Big)f^{n-}_{i-1/2,j}-\Big(|a^1_{i-1/2,j}|-a^1_{i-1/2,j}\Big)f^{n+}_{i-1/2,j},\\ L^n_j(f)=&\Big(|a^2_{i,j+1/2}|-a^2_{i,j+1/2}\Big)f^{n+}_{i,j+1/2}-\Big(|a^2_{i,j+1/2}|+a^2_{i,j+1/2}\Big)f^{n-}_{i,j+1/2}\\ &+\Big(|a^2_{i,j-1/2}|+a^2_{i,j-1/2}\Big)f^{n-}_{i,j-1/2}-\Big(|a^2_{i,j-1/2}|-a^2_{i,j-1/2}\Big)f^{n+}_{i,j-1/2},\end{aligned}$$

with $n=(1)$ and $n=k$ correspond to the solution $f^{(1)}$ and f^k, respectively, at the time-step t^k and analogously for $f^{n\pm}$. Moreover, also the drift is always considered at the time-step t^k. The equations in (2.47) can be rewritten in a compact form with a suitable function $\mathcal{H}$ as follows

$$f^{k+1}_{i,j}=\mathcal{H}(f^k,f^{(1)})+\Delta t\,g^k_{i,j}.\tag{2.48}$$

Now, the KT flux H, given in (2.38), is a combination of the monotonicity preserving Rusanov flux and the monotonicity preserving MUSCL reconstruction. This leads to the SSPRK2-KT scheme to be monotone preserving [**89**]. Thus, $\mathcal{H}$ is a monotone non-decreasing function of its arguments. Then the following discrete entropy inequality holds for the specific Kruzkov entropy pair $(|f|,\mathrm{sgn}(f))$ (see [**123**, Lemma 2.4])

$$|f^{k+1}_{i,j}|\leq|f^k_{i,j}|-\lambda\left(\Psi^{1,k}_{i+1/2,j}-\Psi^{1,k}_{i-1/2,j}+\Psi^{2,k}_{i,j+1/2}-\Psi^{2,k}_{i,j-1/2}\right)+\mathrm{sgn}(f^{k+1})\Delta t\;g^k_{i,j},\tag{2.49}$$

where $\Psi^{1,k}_{\cdot,\cdot}$, $\Psi^{2,k}_{\cdot,\cdot}$ are the conservative entropy fluxes defined for $i,j=1,\dots,N_x$ as follows

$$\begin{aligned}\Psi^{1,k}_{i+1/2,j}=&\frac{H^{x_1,k}_{i+1/2,j}(\max(f^+,0),\max(f^-,0))-H^{x_1,k}_{i+1/2,j}(\min(f^+,0),\min(f^-,0))}{2}\\ &+\frac{H^{x_1,k}_{i+1/2,j}(\max(f^{(1)+},0),\max(f^{(1)-},0))-H^{x_1,k}_{i+1/2,j}(\min(f^{(1)+},0),\min(f^{(1)-},0))}{2},\\ \Psi^{2,k}_{i,j+1/2}=&\frac{H^{x_2,k}_{i,j+1/2}(\max(f^+,0),\max(f^-,0))-H^{x_2,k}_{i,j+1/2}(\min(f^+,0),\min(f^-,0))}{2}\\ &+\frac{H^{x_2,k}_{i,j+1/2}(\max(f^{(1)+},0),\max(f^{(1)-},0))-H^{x_2,k}_{i,j+1/2}(\min(f^{(1)+},0),\min(f^{(1)-},0))}{2}.\end{aligned}$$

Therefore, we have for $k=0,\dots,N_t-1$

$$|f^{k+1}_{i,j}|\leq|f^k_{i,j}|-\lambda\left(\Psi^{1,k}_{i+1/2,j}-\Psi^{1,k}_{i-1/2,j}+\Psi^{2,k}_{i,j+1/2}-\Psi^{2,k}_{i,j-1/2}\right)+\Delta t\;|g^k_{i,j}|.$$

Summing up over all i, j and because of our assumption on f being zero on the boundary, we have

$$\left\|f^{k+1}_{\cdot,\cdot}\right\|_{1,h} \leq \left\|f^{k}_{\cdot,\cdot}\right\|_{1,h} + \Delta t \left\|g^{k}_{\cdot,\cdot}\right\|_{1,h},$$

which iteratively gives us

$$\left\|f^{k+1}_{\cdot,\cdot}\right\|_{1,h} \leq \left\|f^{0}_{\cdot,\cdot}\right\|_{1,h} + \Delta t \sum_{m=0}^{k} \left\|g^{m}_{\cdot,\cdot}\right\|_{1,h}.$$

□

Next, we consider the local consistency error of our SSPRK2-KT at the point (x_1^i, x_2^j, t^k) defined as

$$T^k_{i,j} = \frac{f(x_1^i, x_2^j, t^{k+1}) - f(x_1^i, x_2^j, t^k)}{\Delta t} + \frac{1}{4h}(L^k_i + L^k_j + L^{(1)}_i + L^{(1)}_j)(f(x_1^i, x_2^j, t^k)) - g^k_{i,j}.$$

The accuracy result for the KT scheme, given by Lemma 2.6, the MUSCL reconstruction error given in Equation (60) in [**102**, Section 4.4] for the case when $\kappa = 0$ (in this reference), and the accuracy result for the SSPRK2 scheme from [**77**, Proposition 3.1], give us the following result

LEMMA 2.11. *Let $f \in C^3$ be the exact solution of the Liouville equation* (2.1)*. Under the CFL condition* (2.44)*, the consistency error $T^k_{i,j}$ satisfies the following error estimate*

$$|T^k_{i,j}| = \mathcal{O}(h^2)$$

except possibly at the points of extrema of f where the consistency error can be first-order in h.

Define the error at the point (x_1^i, x_2^j, t^k) as

$$e^k_{i,j} = f^k_{i,j} - f(x_1^i, x_2^j, t^k).$$

Then $e^k_{i,j}$ satisfies (2.47), with the source term given by $-T^k_{i,j}$. Thus, from Lemma 2.10, we obtain

$$\left\|e^{k+1}_{\cdot,\cdot}\right\|_{1,h} \leq \left\|e^{0}_{\cdot,\cdot}\right\|_{1,h} + \Delta t \sum_{m=0}^{k} \left\|T^{m}_{\cdot,\cdot}\right\|_{1,h}.$$

This leads to the following result on the L^1 convergence of the solution obtained using the SSPRK2-KT scheme.

Theorem 2.5. *Let $f \in C^3$ be the exact solution of the Liouville equation* (2.1)*, with finitely many extrema, and let $\left\| f^0_{\cdot,\cdot} - f_0(\cdot,\cdot) \right\|_{1,h} = \mathcal{O}(h^2)$. Under the CFL condition (2.44), the solution $f^k_{i,j}$ obtained with the SSPRK2-KT scheme, given by Algorithm 2.1, is second-order accurate in the discrete L^1-norm as follows*

$$\left\| f^k_{\cdot,\cdot} - f(\cdot,\cdot,t^k) \right\|_{1,h} \leq D(T,\Omega,\lambda)\, h^2, \qquad k = 1,\dots,N_t.$$

The constant $D > 0$ is just depending on its arguments.

2.3.2. Numerical analysis of the Strang splitting scheme

In this section, we deal with the numerical solution of the adjoint equation (2.8) in its form as in (3.32). In this case, we have a terminal condition, and the adjoint problem requires evolution backward in time. For this reason, it is convenient to perform a change of the time variable as follows:

$$s(t) = T - t, \qquad \frac{\partial s}{\partial t} = -1.$$

With this transformation, we can rewrite (3.32) in the following way

$$\partial_s q(x,s) - a(x,s;u(s)) \cdot \nabla q(x,s) = -\theta(x,s), \qquad \text{with} \quad q(x,0) = -\varphi(x)\,. \tag{2.50}$$

Notice that this is (2.8) with $g(x,t) = -\theta(x,t)$ and $q(x,0) = -\varphi(x)$. To solve (2.50), we apply the Strang splitting method [**120**] by first rewriting it as follows

$$\begin{aligned} \partial_s q(x,s) - \operatorname{div}\Big(a(x,s;u(s))\, q(x,s)\Big) + \Big(\operatorname{div} a(x,s;u(s))\Big)\, q(x,s) &= -\theta(x,s), \\ \text{with} \quad q(x,0) &= -\varphi(x)\,, \end{aligned} \tag{2.51}$$

The initial-value problem (2.51) is defined in $\Omega \times [0,T]$. Furthermore, we assume that φ and θ have (by machine precision) compact support for all times inside the interval $[0,T]$. See the numerical experiments section for the specific choices for θ and φ. Then, we solve the problem (2.51) with homogeneous Dirichlet boundary conditions.

We can conveniently illustrate the Strang splitting method applied to (2.51) remaining at the continuous level within one time interval. Let us consider the solution of the adjoint equation (2.51) at time s^k given by $q^k(x)$, $x \in \Omega$. Then, the first step of our solution scheme is to solve the following equation

$$\begin{gathered} \partial_s q(x,s) - \operatorname{div}\Big(a(x,s;u(s))\, q(x,s)\Big) = 0, \\ q(x,s^k) = q^k(x), \qquad s \in [s^k, s^{k+1/2}]. \end{gathered} \tag{2.52}$$

For this purpose, we use the SSPRK2-KT scheme given in Algorithm 2.1. We denote the solution to this problem with q_1.

In the second step, for each x fixed, we solve the following linear ordinary differential equation

$$\begin{aligned} \partial_s q(x,s) &= -\Big(\operatorname{div}(a(x,s;u(s)))\Big)\, q(x,s) - \theta(x,s), \\ q(x,s^k) &= q_1(x,s^{k+1/2}), \qquad s \in [s^k, s^{k+1}]. \end{aligned} \tag{2.53}$$

Let the solution obtained in this step be denoted with q_2.

The last step is to solve (2.52) with the SSPRK2-KT scheme with the initial condition $q_2(\cdot, s^{k+1/2})$ in the time interval $[s^{k+1/2}, s^{k+1}]$. This problem is formulated as follows

$$\begin{aligned} \partial_s q(x,s) - \operatorname{div}\Big(a(x,\tau;u(s))\, q(x,s)\Big) &= 0, \\ q(x,s^{k+1/2}) = q_2(x,s^{k+1}), \qquad s &\in [s^{k+1/2}, s^{k+1}]. \end{aligned} \tag{2.54}$$

In a numerical setting, the solution obtained in this step is the desired solution of the adjoint equation (2.51), and q^{k+1} denotes the adjoint variable at time s^{k+1}. Notice that the value of u in $[s^k, s^{k+1})$ is constant by our numerical approximation strategy for u.

The steps of the Strang splitting scheme are outlined in Algorithm 2.2 below.

Algorithm 2.2 Kurganov-Tadmor-Strang (KTS) scheme

```
Require: $q^0 = -\varphi$, $F$
Ensure: Solve adjoint equation in $q$
Set $k = 0$
while $0 \le k < N_t$ do
    for $1 < i, j < N_x - 1$ do
        Apply one temporal step of Algorithm 2.1 in $(s^k, s^{k+1/2})$, with inputs $q^k_{i,j}$,
            $-F$, $(s^k, s^{k+1/2})$. Denote the solution $q^{k+1/2}_{1,i,j}$.
        In $(s^k, s^{k+1})$, solve (2.53) using exact integration as given in (2.57)
        Apply one temporal step of Algorithm 2.1 in $(s^{k+1/2}, s^{k+1})$, with inputs $q^{k+1}_{2,i,j}$,
            $-F$, $(s^{k+1/2}, s^{k+1})$. Denote the solution with $q^{k+1}_{i,j}$.
    end for
    $k = k + 1$
end while
return $q^k$
```

Now, we discuss some properties of the Strang-splitting scheme described in Algorithm 2.2. For this purpose, we denote with $q^k_{i,j}$ the numerical solution of (2.51) with the generic right-hand side $\mathcal{G}$, at the grid point (x^1_i, x^2_j, s^k).

We have the following discrete L^1 stability estimate.

Lemma 2.12 (Stability of adjoint equation). *Let q be the numerical solution of* (2.51)*, obtained using the KTS scheme, in the interval $[s^k, s^{k+1}]$. Then the following estimate* *holds*

$$\left\|q^{k+1}_{\cdot,\cdot}\right\|_{1,h} \leq \exp(3LT)\left(\left\|q^{0}_{\cdot,\cdot}\right\|_{1,h} + TM\right), \tag{2.55}$$

where $L = \|\operatorname{div} a\|_{L^\infty(\Omega\times[0,T])}$, $M = \|\mathcal{G}\|_{L^\infty_T(L^1(\Omega))}$.

Proof. Let $q^{k+1/2}_{1,\cdot,\cdot}$ be the numerical solution obtained from (2.52). Since (2.52) is solved using the SSPRK2-KT scheme, using the entropy inequality computations as in Lemma 2.10, we have

$$\left\|q^{k+1/2}_{1,\cdot,\cdot}\right\|_{1,h} \leq \left\|q^{k}_{\cdot,\cdot}\right\|_{1,h}. \tag{2.56}$$

Next, denoting the numerical solution as $q^{k+1}_{2,i,j}$ obtained from (2.53), using an integrating factor approach in $[s^k, s^{k+1}]$, we obtain

$$\begin{aligned} q^{k+1}_{2,i,j} &= \exp\left(\mathcal{R}(s^k) - \mathcal{R}(s^{k+1})\right) q^{k+1/2}_{1,i,j} - \exp\left(-\mathcal{R}(s^{k+1})\right)\int_{s^k}^{s^{k+1}} \exp\left(\mathcal{R}(s)\right)\mathcal{G}\, ds \\ &= \Lambda(q_1, \mathcal{G}), \end{aligned} \tag{2.57}$$

where $\mathcal{R} = \int \operatorname{div} a\, ds$.

This equation can be solved exactly, because the drift and $\mathcal{G}$ are given explicitly and the control u is constant in the sub-interval of integration. We exemplify this calculation considering $\mathcal{G}$ being constant in $[s^k, s^{k+1})$ and equal to its value at s^k. Notice that in our case it holds that $\mathcal{R}(s^k) = \left((u^1_2)^{k+1/2} + (u^2_2)^{k+1/2}\right) s^k$.

In general, without any assumptions on the approximation strategy for u and $\mathcal{G}$, but considering the bilinear structure of our drift function (cf. (2.5)) and the assumption on a_0, we can state that there exists an $L > 0$ such that

$$|\operatorname{div}(a(x, s, u))| \leq L, \qquad \forall (x, s) \in \Omega \times [0, T].$$

Thus, we have

$$|\mathcal{R}(s)| \leq LT, \qquad \mathcal{R}(s^k) - \mathcal{R}(s^{k+1}) \leq L\,\Delta t.$$

Hence, by integration we obtain

$$q^{k+1}_{2,i,j} \leq \exp(L\Delta t) q^{k+1/2}_{1,i,j} + \exp(2LT)\int_{s^k}^{s^{k+1}} |\mathcal{G}|\; ds,$$

Further, by using (2.56), we have

$$\begin{aligned} \left\|q^{k+1}_{2,\cdot,\cdot}\right\|_{1,h} &\leq \exp(L\Delta t)\left\|q^{k+1/2}_{1,\cdot,\cdot}\right\|_{1,h} + \exp(2LT)\Delta t M \\ &\leq \exp(L\Delta t)\left\|q^{k}_{\cdot,\cdot}\right\|_{1,h} + \exp(2LT)\Delta t M, \end{aligned} \tag{2.58}$$

where $M = \max_{s\in[0,T]} h^2 \sum_{i,j} |\mathcal{G}(x_i^1, x_j^2, s)|$. Again, since (2.54) is solved using the SSPRK2-KT scheme, we have

$$\begin{aligned}
\left\|q^{k+1}_{\cdot,\cdot}\right\|_{1,h} &\leq \left\|q^{k+1}_{2,\cdot,\cdot}\right\|_{1,h} \\
&\leq \exp(L\Delta t) \left\|q^{k}_{\cdot,\cdot}\right\|_{1,h} + \exp(2LT)\Delta t M \\
&\leq \exp(L\Delta t(k+1)) \left\|q^{0}_{\cdot,\cdot}\right\|_{1,h} + \Delta t M \sum_{m=0}^{k} \exp(L\Delta t m + 2LT) \\
&\leq \exp(L\Delta t N_t) \left\|q^{0}_{\cdot,\cdot}\right\|_{1,h} + N_t \Delta t M \exp(L\Delta t N_t + 2LT) \\
&\leq \exp(3LT) \left(\left\|q^{0}_{\cdot,\cdot}\right\|_{1,h} + TM\right),
\end{aligned}$$

which gives the desired result. □

Next, we consider the local truncation error of our KTS scheme at the point (x_1^i, x_2^j, s^k) defined as [**128**]

$$Z^k_{i,j} = q(x_i^1, x_j^2, s^{k+1}) - \left[\mathcal{H}(q_2, q_2^{(1)}) \circ \Lambda(q_1, \mathcal{G}) \circ \mathcal{H}(q^k, q^{(1)})\right] (q(x_i^1, x_j^2, s^k)),$$

where $\mathcal{H}$ is the SSPRK2-KT operator given in (2.48) and Λ is the exact integration operator for (2.53) at time s^k, defined in (2.57). We have the following temporal error estimate for the continuous Strang splitting scheme (for its proof see [**120**, Page 510], [**118**, Eq. (1.7)],[**45**, Eq. (2.13)]).

Lemma 2.13 (Time error Strang-splitting). *Let $S = S(\Delta t)$ be the exact solution operator of* (2.51) *in $[s^k, s^{k+1}]$, i.e., $S\,q^k = q^{k+1}$. Denote with q_{SP} the solution of* (2.51) *with the Strang splitting scheme, given by* (2.52)–(2.54)*, applied at the continuous level (no discretization of the spatial and the temporal operators) in the time interval $[s^k, s^{k+1}]$ and with a smooth initial condition $\bar{q}(\cdot, s^k)$. This solution can be written as follows*

$$q_{SP}(\cdot, s^{k+1}) = \left(S_2 \circ \Lambda \circ S_1\right) \bar{q}(\cdot, s^k),$$

where $S_1 = S_1(\Delta t)$ denotes the exact integration of $\partial_s q - \operatorname{div}(aq) = 0$ in time interval $[s^k, s^{k+1/2}]$, and $S_2 = S_2(\Delta t)$ the same operator for $[s^{k+1/2}, s^k]$. Then the following error estimate holds

$$\max_{x\in\Omega} \left| q_{SP}(x, s^{k+1}) - S\,\bar{q}(x, s^{k+1}) \right| = \mathcal{O}(\Delta t^3). \tag{2.59}$$

With this result and the truncation error estimate of the SSPRK2-KT scheme given in (2.11), we have the following result

Lemma 2.14. *Let $q \in C^3$ be the exact solution of the adjoint equation* (2.51) *Under the CFL condition* (2.44)*, the truncation error $Z^k_{i,j}$ satisfies the following error estimate*

$$|Z^k_{i,j}| = \mathcal{O}(h^3)$$

except possibly at the points of extrema of the exact solution $q(x,t)$.

Define the error at the point (x_1^i, x_2^j, s^k) as

$$e_{i,j}^k = q_{i,j}^k - q(x_1^i, x_2^j, s^k).$$

Then $e_{i,j}^k$ satisfies (2.51), with the right-hand side being $\dfrac{Z_{i,j}^k}{\Delta t}$. Thus, from Lemma 2.12 we obtain

$$\left\| e_{\cdot,\cdot}^{k+1} \right\|_{1,h} \leq \exp(LT) \left(\left\| e_{\cdot,\cdot}^{0} \right\|_{1,h} + \frac{MT}{\Delta t} \right),$$

where $M = \max_{k\in\{0,\ldots,N_t\}} h^2 \sum_{i,j} |Z_{i,j}^k|$. This leads to the following result on the L^1 convergence of the solution obtained using the KTS scheme.

THEOREM 2.6. *Let $q \in C^3$ be the exact solution of the adjoint equation* (2.51)*, with countably many extrema, and let $\left\| q_{\cdot,\cdot}^0 + \varphi(\cdot,\cdot) \right\|_{1,h} = \mathcal{O}(h^2)$. Under the CFL condition* (2.44)*, the solution $q_{i,j}^k$ obtained with the KTS scheme, given by Algorithm 2.2, is second-order accurate in the discrete L^1-norm as follows*

$$\left\| q_{\cdot,\cdot}^k - q(\cdot,\cdot,t^k) \right\|_{1,h} \leq E(T,\Omega,\lambda)\, h^2 \qquad\qquad k = 1,\ldots,N_t.$$

The constant $E > 0$ is just depending on its arguments.

REMARK 2.6. *We remark that results similar to Theorem 2.6 have been obtained in* [**45, 46**]*. However, in these papers the equation that has been considered is the convection diffusion equation, which is parabolic, whereas we have a hyperbolic transport (adjoint) equation with a source term. Furthermore, we employ a different analysis using an entropy inequality technique for proving the discrete stability estimate that is subsequently used for proving the convergence error estimate.*

2.3.3. Verification of the implementation

In this section, we present results of numerical experiments to validate the accuracy of our numerical framework. We have proved second-order accuracy of our SSPRK2-KT scheme for the Liouville equation in Theorem 2.5. In order to validate this estimate, we define a setting that admits a known exact solution. For this purpose, we choose the following control function

$$u(t) = \begin{pmatrix} 0.05\,t & 0.002 \\ 0.5 & -0.001 \end{pmatrix},$$

which results in the following drift

$$a(x,t) = \begin{pmatrix} 0.05\,t \\ 0.5 \end{pmatrix} + \begin{pmatrix} 0.002\,x_1 \\ -0.001\,x_2 \end{pmatrix}. \tag{2.60}$$

Further, we take the initial condition

$$f_0(x) = \frac{1}{2\pi\bar{\sigma}_0} \exp\left(-\frac{1}{2}\left[\frac{x_1^2}{\bar{\sigma}_0} + \frac{x_2^2}{\bar{\sigma}_0}\right]\right), \tag{2.61}$$

where $\bar{\sigma}_0 = \frac{1}{4}$.

With this setting, the Liouville problem

$$\partial_t f + \operatorname{div}(a\, f) = 0, \qquad \text{with} \quad f_{|t=0} = f_0,$$

admits the solution

$$\bar{f}(x,t) = \frac{1}{2\pi\sqrt{\bar{\sigma}_1(t)\bar{\sigma}_2(t)}} \exp\left(-\frac{1}{2}\left[\frac{(x_1 - \bar{m}_1(t))^2}{\bar{\sigma}_1(t)} + \frac{(x_2 - \bar{m}_2(t))^2}{\bar{\sigma}_2(t)}\right]\right), \tag{2.62}$$

In (2.62) the mean $\bar{m}(t) = (\bar{m}_1(t), \bar{m}_2(t))$ and the variance $\bar{\sigma} = (\bar{\sigma}_1(t), \bar{\sigma}_2(t))$ are the solutions to (2.7) with the initial conditions $\bar{m}_0 = (0, 0)$ and $\bar{\sigma}_0 = (1, 1)$. Now, we use this setting to determine the solution error of our algorithm. For this purpose, we solve the corresponding Liouville problem and report the values of the discrete L^1 norm of the solution error given by

$$e_{KT}(f_h) \coloneqq \left\| f_h(\cdot, T) - \bar{f}(\cdot, T) \right\|_{1,h}.$$

In Table 2.1, the values of e_{KT} corresponding to different grids are presented, and in Figure 2.1, we compare the rate of change of these values with that of first- and second-order accuracies. We see that the obtained numerical accuracy lies between these reference rates, becoming closer to second-order by refining the mesh size.

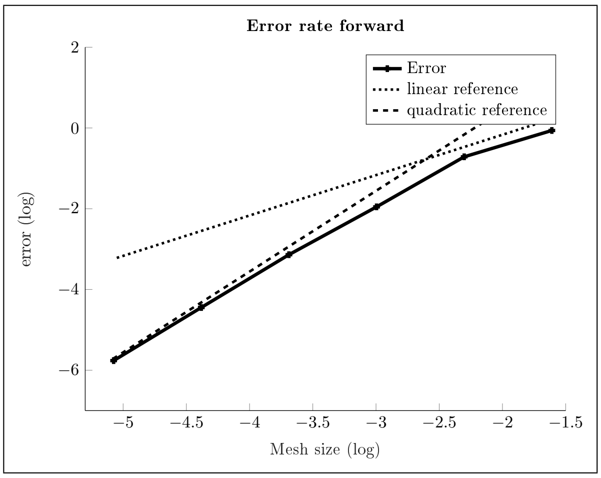


Figure 2.1. Logarithmic plot of accuracy test for the SSPRK2-KT scheme.

N_x	N_t	$e_{KT}(f_h)$
5	20	0.9399
$2 \cdot 5$	$2 \cdot 20$	0.4897
$2^2 \cdot 5$	$2^2 \cdot 20$	0.1417
$2^3 \cdot 5$	$2^3 \cdot 20$	0.0433
$2^4 \cdot 5$	$2^4 \cdot 20$	0.0117
$2^5 \cdot 5$	$2^5 \cdot 20$	0.0031

Table 2.1. L^1-norm of solution error for the SSPRK2-KT scheme.

Next, we validate our estimate for the KTS scheme in solving a transport problem with source term (the adjoint problem) as given in Theorem 2.6. We proceed in a similar way as for the Liouville problem. In fact, since (2.62) solves the Liouville problem with the drift (2.60) and the initial condition (2.61), it is easy to verify that this solution satisfies the problem

$$\partial_t q - \tilde{a}(x,t) \cdot \nabla q = -\theta, \qquad q(0) = -\varphi,$$

where $\tilde{a}(x,t) = -a(x,t)$, $\theta = \bar{f}\nabla a$ and $-\varphi = f_0$. Thus, we have the solution $\bar{q} = \bar{f}$. However, notice that the KTS scheme uses the Strang splitting in order to accommodate the source term $-\theta$. Therefore, the solution $\bar{q} = \bar{f}$ is appropriate to independently test the KTS scheme. Thus, we define

$$e_{KTS}(q_h) = \|q_h(\cdot, T) - \bar{q}(T)\|_{1,h} \,.$$

Hence, we perform a second series of experiments where we compute the values of this norm in correspondence to solutions obtained on different grids. These values are reported in Table 2.2, and in Figure 2.2, we compare the rate of change of e_{KTS} with that of first- and second-order accuracies. Also in this case, we see that the resulting rate of convergence is approximately of second-order.

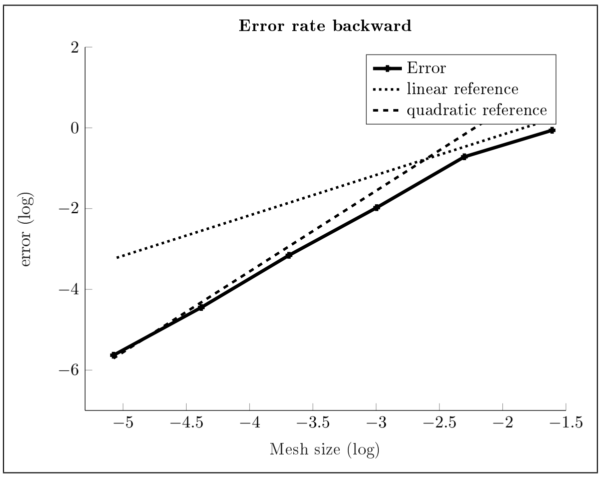


Figure 2.2. Logarithmic plot of accuracy test for the KTS scheme.

N_x	N_t	$e_{KTS}(q_h)$
5	20	0.9414
$2 \cdot 5$	$2 \cdot 20$	0.4884
$2^2 \cdot 5$	$2^2 \cdot 20$	0.1385
$2^3 \cdot 5$	$2^3 \cdot 20$	0.0425
$2^4 \cdot 5$	$2^4 \cdot 20$	0.0116
$2^5 \cdot 5$	$2^5 \cdot 20$	0.0035

Table 2.2. L^1-norm of solution error for the KTS scheme.

After analyzing both the theory and the implementation of our solvers for the Liouville equation and its adjoint, we demonstrate in detail how to solve optimal control problems governed by the Liouville equation in the next chapter.

Chapter 3

Ensemble control problems

The notion of *ensemble control* was proposed by R.W. Brockett in [**32**], and further in [**33, 34**], while considering the problem of a trade-off between the complexity of implementing a control strategy and the performance of the control system. For the former, Brockett discusses the concept of minimum attention control that results in costs of the control that involve a time derivative of the control function. For the latter, he emphasizes the advantage of considering an ensemble of trajectories, which stem from a distribution of initial conditions, rather than individual trajectories. By these two considerations, Brockett concludes that the natural setting for investigating both aspects of the resulting control problem is by means of the Liouville (or continuity) equation that governs the evolution of the ensemble of trajectories.

Therefore, the problem of controlling a trajectory of a finite-dimensional dynamical system is lifted to the problem of controlling a continuum of dynamical systems with the same control strategy. Specifically, this setting results in the problem of determining a single closed- or open-loop controller, which applies to a particular system over an infinite number of repeated trials, or to steer a family of finite-dimensional dynamical systems. As discussed by Brockett, this approach represents a new control framework that is able to address a number of issues as uncertainty in initial conditions and the trade-off mentioned above.

We follow a standard scheme. First, in Section 3.2 we define the Liouville control-to-state map G, namely the map that associates to any control u the unique solution $f = G(u)$, called state, to the corresponding Liouville equation, and study its main properties. A fundamental issue in this part is to show Fréchet differentiability of G in a suitable topology. Our method to prove this property relies on performing stability estimates on the Liouville equation (see Section 3.2.2). Now, dealing with the growth in space of our drift function at $+\infty$ requires the use of weighted norms; moreover, due to the hyperbolicity of transport and continuity equations, a loss of regularity occurs, which requires to consider both higher smoothness and higher integrability on the initial data. Namely, the initial data is assumed to be an element of H_k^m with both $m \geq 2$ and $k \geq 2$.

In Section 3.3, we complete the investigation of the ensemble optimal control problem in the case of attracting L^2-integrable potentials; the adaptations needed to treat the case of quadratic potentials are mentioned in Section 3.3.4. The first step consists in establishing the existence of optimal controls (see Theorem 3.2). Then, we characterize these optimal controls as solutions of a first-order optimality system. This system can be interpreted in terms of the Fréchet differential of the reduced functional $J_r(u) := J\Big(G(u), u\Big)$ set to zero; the reduced functional is defined below. We remark that the differentiability properties of J (and J_r) change radically depending on the choice of the optimization weights. For instance, if we only consider L^2 control costs, then the optimization space is $L^2(0,T)$ and we have Fréchet differentiability of the cost functional. This is the standard case. If we additionally include L^1 control costs, then we have a semi-smooth optimal control problem and we have to resort to the use of sub-differentials. Finally, if we additionally consider H^1 control costs, then $H^1(0,T)$ is the appropriate control space, and the optimality condition accounts for this fact. If all weights are positive and taking into account control constraints, we have an optimal control problem whose structure (to the best of our knowledge) has never been investigated in PDE optimization. For this general case, we prove existence of Lagrange multipliers (see Theorem 3.3) and derive the optimality system.

In Section 3.3.3, we address the uniqueness of optimal ensemble controls, in the special case of only L^2 control costs. More precisely, in Theorem 3.4 we show uniqueness of optimal controls for the control-constrained problem, provided a smallness condition is satisfied; such a condition requires the time T and the size of the data f_0, g, θ and φ to be small enough, or the coefficient of the L^2 control cost to be sufficiently large. This part of the analysis exploits in a fundamental way the optimality system that we derive in Section 3.3.2, and the characterisation of optimal controls as solutions to it.

In Section 3.4, we illustrate our numerical strategy to solve the optimal control problems governed by the Liouville equation. For this goal, we use a projected semi-smooth Krylov-Newton method. Notice that, despite Newton methods are well-known and also their expansion to semi-smooth functions is used frequently, the version including a H^1-projection is not prevailing. We validate our implementation and perform several experiments that demonstrate the ability of our optimal control to achieve the given tasks. Moreover, we elucidate that our framework can be extended to more general problems.

Now, let us recall the control mechanism. The focus of ensemble control is the development of a control strategy for the differential model (2.2) augmented with a control mechanism, as follows:

$$\dot{\xi}(t) = a(\xi(t), t; u), \tag{3.1}$$

where u denotes the control function. We refer to [**33, 34**] for a discussion on the choice of u as a function of time only, which corresponds to a so-called open-loop control, or as a function of time and of the state variable, which may represent a feedback law. In this work, while considering the controlled Liouville (2.6) model in a general setting that accommodates both choices, we focus our attention on open-loop optimal control problems. This is motivated by the fact that the most used control mechanisms for (3.1) are the linear and bilinear ones. We choose

$$a(x, t; u) \;=\; a_0(x, t) \;+\; a_1 u_1(t) \;+\; x \circ a_2 u_2(t)\,,$$

where a_0 is a given smooth vector field and $a_1, a_2 \in \mathbb{R}$ are given constants and $u \;=\; (u_1, u_2)$ is the control.

In the next section, we illustrate the formulation of Liouville ensemble optimal control problems and discuss the chosen control mechanism and the constitutive terms of an ensemble cost functional.

3.1. Formulation of ensemble optimal control problems

In order to discuss Brockett's formulation of ensemble control, consider the following ODE optimal control problem:

$$\min\; j(\xi, u) \;:=\; \int_0^T \Big(\theta\big(\xi(t)\big) \,+\, \kappa\big(u(t)\big)\Big)\, dt \,+\, \varphi\big(\xi(T)\big) \tag{3.2}$$

$$\text{s.t.} \qquad \dot{\xi}(t) \;=\; a\big(\xi(t), t; u(t)\big)\,, \qquad \xi(0) \;=\; \xi_0\,, \tag{3.3}$$

The functions θ, κ and φ are usually taken to be continuous convex functions of their arguments; we will better specify their properties later on.

The optimal control function u is sought in the following set of admissible controls

$$U_{ad} \;:=\; \Big\{ u \,\in\, \mathbb{L}^\infty_T(\mathbb{R}^d) \;\Big|\; u^a \,\le\, u(t) \,\le\, u^b \qquad \text{for a.e. } \; t \,\in\, [0, T] \Big\}\;. \tag{3.4}$$

In particular, in the case of (2.5), we have two box constraints $u^a \;=\; (u_1^a, u_2^a)$ and $u^b \;=\; (u_1^b, u_2^b)$, where $u_\iota^a \,<\, u_\iota^b$, $\iota \,=\, 1, 2$, are given vectors in $\mathbb{R}^d$. We remark that, if we include H^1 control costs, the resulting u is continuous because of the compact embedding $H^1(0, T) \subset\subset C([0, T])$. Clearly, the optimal control function u that solves (3.2)–(3.3) with $u \in U_{ad}$ depends on the fixed initial condition ξ_0. Furthermore, it represents a control strategy that is determined once and for all times for the given ξ_0

and the given optimization setting. Therefore, no uncertainty on the initial condition is taken into account in the formulation (3.2)–(3.3). Hence, from this point of view, the resulting control is not robust. On the other hand, a closed loop control, say, $u = u(x,t)$, would appropriately control the system based on the actual state of the system. However, as pointed out in [**33**], the cost of implementing such a control mechanism is often prohibitive and may be not justified by real applications.

With the purpose to strike a balance between the desired performance of the system and the cost of implementing an effective control, the ensemble control strategy considers instead a density of initial conditions, and therefore ensemble of trajectories. In this way, it aims at achieving robustness, while choosing control costs which promote controls allowing for easier implementation.

Thus, one is led to the formulation of the following ensemble optimal control problem:

$$\min_{u\in U_{ad}} J(f,u) := \int_0^T \int_{\mathbb{R}^d} \theta(x)\, f(x,t)\, dx\, dt + \int_{\mathbb{R}^d} \varphi(x)\, f(x,T)\, dx + \int_0^T \kappa\Big(u(t)\Big)\, dt \tag{3.5}$$

$$\text{s.t.} \qquad \partial_t f + \operatorname{div}\Big(a(x,t;u)\, f\Big) = 0\,, \qquad f_{|t=0} = f_0\,. \tag{3.6}$$

This problem is defined on the space-time cylinder $\mathbb{R}^d \times [0,T]$, for some $T > 0$ fixed. In this formulation, the initial density f_0 represents the probability distribution of the initial condition ξ_0 in (3.2)–(3.3), and thus it models the known uncertainty on the initial data.

Next, we discuss some specific choices of the optimization components in (3.2)–(3.3), and correspondingly in (3.5)–(3.6).

For example, if $\xi = 0$ is a critical point for (3.3), which requires $a(0,t;u) = 0$, then the choice $\theta(x) = x^2$ appears standard for stabilization purposes. Usually, in this context, the so-called L^2 cost of the control is considered, which corresponds to the choice $\kappa(u) = \gamma\, u^2$, where $\gamma > 0$ is the weight of the cost of the control. On the other hand, if the purpose of the control in (3.2)–(3.3) is to track a desired and even non-attainable trajectory $\xi_D \in L^2(0,T;\mathbb{R}^d)$, and to come close to a given final configuration $\xi_T \in \mathbb{R}^d$ at the final time (possibly with $\xi_D(T) \neq \xi_T$), then a natural choice appears to be $\theta\Big(x(t)\Big) = \alpha\Big(x(t) - \xi_D(t)\Big)^2$ and $\varphi\Big(x(T)\Big) = \beta\Big(x(T) - \xi_T\Big)^2$, with appropriately chosen weights α, $\beta > 0$. However, in the context of ensemble control, as in (3.5), the choice of θ and φ as convex functions is problematic because of integrability issues. On the other hand, we remark that the role of these functions is to define attracting potentials, that is, to define a well or sink centred at a minimum point such that the minus gradient of the potential is directed towards this minimum. For this purpose, a possible choice is also $\theta(x) = 1 - \exp(-x^2)$, with the minimum at $x = 0$. In our analysis, we are able to address both cases in the framework of weighted Sobolev spaces. Specifically, the case of attracting potentials θ and φ which are both L^2 integrable, and the case of θ and φ which are quadratic functions. Notice that, in any case, the modelling choice for (3.2)–(3.3) translates without changes to (3.5)–(3.6).

As discussed in [**34, 33, 32**], the choice of the cost function κ should be such that the effort of implementing the control strategy is as small as possible. In this sense, the cost of a slowly varying control function, and a control that does not act for all times, should be smaller than the cost corresponding to a control that has large variations and acts for all times. From this perspective, a constant input that controls the system is the cheapest choice, and the next possible choice is a control that slowly changes in time. This requirement leads naturally to a cost of the form

$$\nu \int_0^T \left(\frac{du}{dt}(t) \right)^2 dt,$$

with $\nu \geq 0$. In fact, as ν is taken larger, the resulting optimal control will have smaller values of its time derivative, that is, a slowly varying control, which is called "minimum attention control" in [**32**].

More recently, there has been a surge of interest in L^1-costs, originating from signal reconstruction and magnetic resonance imaging [**38**]. This cost is given by

$$\delta \int_0^T |u(t)| \, dt,$$

where $\delta \geq 0$. The effect of this cost is that it promotes sparsity of the control function, in the sense that, as $\delta > 0$ is increased, the u resulting from the minimisation procedure will be zero on open intervals in $]0, T[$, and these intervals become larger and eventually cover all of $]0, T[$ as $\delta \to +\infty$. In the present chapter, we introduce the L^1-cost in the context of ensemble control and call the resulting sparse control a "minimum action control".

All together, we specify the term $\int_0^T \kappa\big(u(t)\big)\, dt$ in (3.2) and in (3.5) as follows:

$$\kappa\big(u(t)\big) := \frac{\gamma}{2} \big(u(t)\big)^2 + \delta \, |u(t)| + \frac{\nu}{2} \left(\frac{du}{dt}(t) \right)^2, \tag{3.7}$$

where $\gamma + \delta + \nu > 0$, γ, δ, $\nu \geq 0$ and the factor $1/2$ is chosen for convenience of later calculations. Notice that different choices of the value of the positive coefficients γ, δ, ν will result in different features of the resulting optimal control function.

3.2. The Liouville control-to-state map

In this section, we define the Liouville control-to-state map and investigate its continuity and differentiability properties. For reasons which will appear clear in the following analysis, we need to resort to weighted spaces H^m_k, as introduced in Section 2.2.2. We start by making a remark.

Remark 3.1. *Throughout this section, the data of the Liouville equation has to be thought as fixed. Specifically, for $m \in \mathbb{N}$ and $k \in \mathbb{N}$, we take an initial datum $f_0 \in H^m_k$, a source term $g \in L^1_T(H^m_k)$, and a drift function $a_0 \in L^1_T(C^{m+1})$, with $\nabla a_0 \in L^1_T(C^m_b)$.*

We are then interested in the dependence of the solution f to the Liouville equation (2.9)*, with drift a given by* (2.5)*, on the control state $u \in U_{ad}$, where U_{ad} has been defined in* (3.4)*.*

3.2.1. Definition and continuity properties

We remark that the statements of Theorems 2.2 and 2.4 cover the case of the Liouville equation with the controlled drift function given by (2.5), where $u \in U_{ad}$. In particular, the next proposition-definition immediately follows.

Proposition 3.1. *For fixed data f_0, g and a_0 as in Remark 3.1, let us consider drift functions a of the form* (2.5)*, with $u \in U_{ad}$. We introduce the* Liouville control-to-state map *G, defined by*

$$G : U_{ad} \to L^\infty\Big([0,T]; L^2(\mathbb{R}^d)\Big), \qquad u \mapsto f := G(u),$$

where f is the unique solution to the Liouville equation with the given data.

Then G is well-defined.

Let us make an important comment about the previous definition.

Remark 3.2. *The theory developed in Sections 2.2.1 and 2.2.2 entails that the solution f actually belongs to $C_T(H^m_k)$. However, due to a* loss of regularity*, both in m and k, when proving Fréchet differentiability of G, it is convenient to look at G as a map with values in the space with the weakest topology.*

Finally, we consider L^∞ regularity with respect to time, because it will be convenient also to look at weak continuity properties of G, see Proposition 3.2 below.

Next, we study some properties of the map G that are relevant for the analysis of ensemble optimal control problems. We start by establishing that G is weak-weak continuous from U_{ad} into $L^\infty_T(L^2)$. Notice that we do not need any restriction on m and k in this case.

Proposition 3.2. *Take $m \geq 0$ and $k \geq 0$ and initial data f_0, g and a_0 as in Remark 3.1. Let $u \in U_{ad}$ and $\left(u^l\right)_l \subset U_{ad}$ be a sequence of controls, and assume that $u^l \overset{*}{\rightharpoonup} u$ in $\mathbb{L}^\infty_T$.*
Then $G(u^l) \overset{*}{\rightharpoonup} G(u)$ in the weak-$*$ topology of $L^\infty_T(L^2)$.

Proof. Of course, it is enough to prove the previous proposition in the case of minimal regularity and integrability, namely for $m = k = 0$.

By definition of the set U_{ad}, we infer that $(u^l)_l$ is uniformly bounded in $\mathbb{L}^\infty_T$. On the other hand, by hypotheses and Theorem 2.2, for all $l \in \mathbb{N}$ there exists a unique $f^l := G(u^l) \in C_T(L^2)$ which solves the Liouville equation (2.9). In addition, by inequality (2.12), we deduce that $\left(f^l\right)_l$ is uniformly bounded in $C_T(L^2)$. Then there exists $f \in L^\infty_T(L^2)$ such that, up to extraction of a subsequence, $f^l \overset{*}{\rightharpoonup} f$ in $L^\infty_T(L^2)$. So, the proof reduces to showing that f is a weak solution to the Liouville equation

$$\partial_t f \, + \, \mathrm{div}\left(a(t,x;u)\, f\right) \, = \, g\,, \qquad \text{with} \qquad f_{|t=0} \, = \, f_0\,. \tag{3.8}$$

Indeed, if this is the case, by uniqueness we get $f \, = \, G(u)$ and that the whole sequence $\left(f^l\right)_l$ converges.

The previous property follows by passing to the limit in the weak formulation of the equation for f^l. This can be easily obtained Notice that, in order to treat the products between f^l and u^l, one also needs to establish strong convergence for f^l in suitable spaces (as done in the proof to Proposition 2.1).

In order to prove our claim, we need to pass to the limit in the weak formulation of the Liouville equation for f^l, when $l \to +\infty$. Recalling also our special choice (2.5), it is easy to see that the only term which presents some difficulty is the non-linear term

$$\int_0^T\!\!\int_{\mathbb{R}^d} f^l\, \left(u^l_1(t) + x \circ u^l_2(t)\right) \cdot \nabla\phi \, dx\, dt\,, \quad \text{for any fixed} \quad \phi \in C^\infty_c\left(\mathbb{R}^d \times [0,T[\right). \tag{3.9}$$

Therefore, let us focus on the convergence of this integral. First, by inspection of the equation $\partial_t f^l \, = \, -\,\mathrm{div}\left(a(x,t;u^l)\, f^l\right) + g$, we discover that $\left(\partial_t f^l\right)_l \subset L^1_T(H^{-1}_{\rm loc})$, which implies that $\left(f^l\right)_l \subset W^{1,1}_T(H^{-1}_{\rm loc})$. Then, by the Rellich-Kondrachov theorem and Cantor's diagonal procedure, we discover that, up to an extraction of a subsequence that we do not relabel, $\left(f^l\right)_l$ is compact, and then strongly convergent, in $L^1_T(H^{-2}_{\rm loc})$. Interpolating this compactness result with the uniform boundedness in $L^\infty_T(L^2_{\rm loc})$, we discover that $f^l \to f$ in $L^1_T(H^{-s}_{\rm loc})$, for any $s \in (0,2)$; see [**124**]. In view of the uniform boundedness of $\left(u^l_1\right)_l$ and $\left(u^l_2\right)_l$ in L^∞_T, the previous property is enough to pass to the limit in the integral (3.9), and prove that it converges to

$$\int_0^T \int_{\mathbb{R}^d} f\, \left(u_1(t) \, + \, x \circ u_2(t)\right) \cdot \nabla\phi \, dx\, dt\,,$$

for all given $\phi \in C^\infty_c\Big(\mathbb{R}^d \times [0,T[\Big)$. Thus, we get that (3.8) is satisfied, and then we can conclude the proof as already mentioned above. □

For the analysis of our optimal control problem, we need stronger regularity properties for G. We start by showing Lipschitz continuity, which will be the basis to prove Gâteaux differentiability of G. The key here is to perform careful estimates in order to identify the correct topology. The reason is that, due to hyperbolicity of the Liouville equation, stability estimates involve a loss of regularity.

Lemma 3.1. *Let the data f_0, g and a_0 be fixed as in Remark 3.1 above, with $m \geq 1$ and $k \geq 1$. Let u and v be in U_{ad}, and denote by $G(u)$ and $G(v)$ the corresponding $C_T(H^m_k)$ solutions to (2.9), with drift a given by (2.5). Set $\delta G := G(u) - G(v)$. Then there exists a constant $C > 0$, independent of the data and respective solutions, such that, for all $1 \leq \ell \leq k$, if we set*

(3.10)
$$K^{(\ell)}_0 := C \exp\Big(C\left(\|\nabla a_0\|_{L^1_T(C^1_b)} + \|u\|_{\mathbb{L}^1_T} + \|v\|_{\mathbb{L}^1_T}\right)\Big)\left(\|f_0\|_{H^1_\ell} + \|g\|_{L^1_T(H^1_\ell)}\right),$$

then, for all $t \in [0,T]$, one has

$$\|\delta G(t)\|_{L^2_{\ell-1}} \leq K^{(\ell)}_0 \int_0^t |u(s) - v(s)|\, ds\,.$$

If moreover $m \geq 2$ and we set

(3.11)
$$K^{(\ell)}_1 := C \exp\Big(C\left(\|\nabla a_0\|_{L^1_T(C^2_b)} + \|u\|_{\mathbb{L}^1_T} + \|v\|_{\mathbb{L}^1_T}\right)\Big)\left(\|f_0\|_{H^2_\ell} + \|g\|_{L^1_T(H^2_\ell)}\right),$$

we also have

$$\|\delta G(t)\|_{H^1_{\ell-1}} \leq K^{(\ell)}_1 \int_0^t |u(s) - v(s)|\, ds\,.$$

Proof. By linearity of the Liouville equation, we find that δG satisfies

$$\partial_t \delta G + \operatorname{div}\Big(a(t,x;u)\,\delta G\Big) = -\operatorname{div}\Big(\overline{a}(x,t;u-v)\,G(v)\Big)\,, \qquad \delta G_{|t=0} = 0\,, \tag{3.12}$$

where we have set $\overline{a}(t,x;u-v) := a(t,x;u) - a(t,x;v) = (u_1 - v_1) + x \circ (u_2 - v_2)$. Applying $L^2_{\ell-1}$ estimates of Theorem 2.2 to equation (3.12), we immediately get

$$\|\delta G(t)\|_{L^2_{\ell-1}} \leq C \exp\left(C\int_0^t \|\nabla a(s,x;u)\|_{L^\infty}\, ds\right) \int_0^t \left\|\operatorname{div}\Big(\overline{a}(s,x;u-v)\,G(v)\Big)\right\|_{L^2_{\ell-1}} ds\,.$$

By explicit computations and using the Leibniz rule, we deduce that

(3.13)
$$\begin{aligned}\left\|\operatorname{div}\Big(\overline{a}(s,x;u-v)\,G(v)\Big)\right\|_{L^2_{\ell-1}} &\leq |u(s) - v(s)| \left(\|G(v)\|_{L^2_{\ell-1}} + \|\nabla G(v)\|_{L^2_\ell}\right) \\ &\leq C\,|u(s) - v(s)| \exp\left(C\int_0^s \|\nabla a(s,x;v)\|_{C^1_b}\, ds\right)\left(\|f_0\|_{H^1_\ell} + \int_0^s \|g(s)\|_{H^1_\ell}\, ds\right),\end{aligned}$$

where the second inequality holds true in view of the bound $\|G(v)\|_{L^2_{\ell-1}} + \|\nabla G(v)\|_{L^2_\ell} \leq \|G(v)\|_{H^1_\ell}$ and Lemma 2.5. This estimate completes the proof of the first inequality.

Now, we focus on $H^1_{\ell-1}$ bounds for δG. Thanks to Lemma 2.5, we have

(3.14)
$$\|\delta G(t)\|_{H^1_{\ell-1}} \leq C \exp\left(C \int_0^t \|\nabla a(s,x;u)\|_{C^1_b}\, ds\right) \int_0^t \left\|\mathrm{div}\left(\overline{a}(s,x;u-v)G(v)\right)\right\|_{H^1_{\ell-1}} ds\,.$$

By definition, we have that $\|f\|_{H^1_{\ell-1}} = \|f\|_{L^2_{\ell-1}} + \|\nabla f\|_{L^2_{\ell-1}}$. Concerning the first term, we have

(3.15)
$$\begin{aligned}\left\|\mathrm{div}\left(\overline{a}(s,x;u-v)\,G(v)\right)\right\|_{L^2_{\ell-1}} &= \|\mathrm{div}\,\overline{a}\,G(v)\|_{L^2_{\ell-1}} + \|\overline{a}\cdot\nabla G(v)\|_{L^2_{\ell-1}} \\ &\leq C\,|u(s)-v(s)|\,\left(\|G(v)\|_{L^2_{\ell-1}} + \|\nabla G(v)\|_{L^2_\ell}\right) \\ &\leq C\,|u(s)-v(s)|\;\|G(v)\|_{H^1_\ell}\,.\end{aligned}$$

Next, we need to bound in $L^2_{\ell-1}$ the quantity $\nabla\,\mathrm{div}\left(\overline{a}(s,x;u-v)\,G(v)\right)$. For this purpose, we have to control four terms. Notice that $\nabla\,\mathrm{div}\,\overline{a} \equiv 0$. Moreover, we can write

$$\|\mathrm{div}\,\overline{a}\;\nabla G(v)\|_{L^2_{\ell-1}} \leq |u(s)-v(s)|\;\|\nabla G(v)\|_{L^2_{\ell-1}}\,, \tag{3.16}$$

and the same estimate holds true also for the term $\nabla\overline{a}\cdot\nabla G(v)$. Finally, we have

$$\left\|\overline{a}\cdot\nabla^2 G(v)\right\|_{L^2_{\ell-1}} \leq C\,|u(s)-v(s)|\;\left\|\nabla^2 G(v)\right\|_{L^2_\ell}\,. \tag{3.17}$$

Putting (3.15), (3.16) and (3.17) together, we infer the control

$$\left\|\mathrm{div}\left(\overline{a}(s,x;u-v)\,G(v)\right)\right\|_{H^1_{\ell-1}} \leq C\,|u(s)\,-\,v(s)|\;\|G(v)\|_{H^2_\ell}\,.$$

Inserting this last inequality into (3.14) and using the bounds of Theorem 2.4, we finally get the claimed estimate for the H^1-type norms of δG. □

3.2.2. Differentiability of the control-to-state map

In this section, we investigate differentiability properties of the control-to-state map G.

With Lemma 3.1 at hand, we can establish Gâteaux differentiability of G. For any given u in an open set $U_0 \subset U_{ad}$, let $G(u)$ be the corresponding solution to the Liouville equation, as defined in Proposition 3.1, and let $\delta u = (\delta u_1, \delta u_2)$ be an admissible variation of u, such that $u + \varepsilon\delta u \in U_{ad}$ for $\varepsilon \in \mathbb{R}\setminus\{0\}$ sufficiently small. Then the Gâteaux derivative of G with respect to the variation δu at u is defined as the limit

(whenever such a limit exists)

$$\delta_{\delta u}G(u) \;:=\; \lim_{\varepsilon\to 0}\frac{G(u+\varepsilon\delta u)\;-\;G(u)}{\varepsilon}\,. \tag{3.18}$$

The next proposition holds true.

Proposition 3.3. *Let $m \geq 2$ and $k \geq 2$. Let the data f_0, g and a_0 be fixed as in Remark 3.1 above. Let u belong to $\operatorname{int} U_{ad}$, where $\operatorname{int} U_{ad}$ denotes the interior part of the set U_{ad}.*

Then, for any admissible variation δu of u, the limit (3.18) exists in $L^\infty_T(L^2)$. In particular, the control-to-state map G is Gâteaux differentiable at u. Moreover, $\delta_{\delta u}G$ satisfies the Liouville problem

$$\partial_t\delta_{\delta u}G\;+\;\operatorname{div}\Big(a(t,x;u)\,\delta_{\delta u}G\Big)\;=\;-\;\operatorname{div}\Big(\overline{a}(t,x;\delta u)\,G(u)\Big)\,,\qquad \delta_{\delta u}G_{|t=0}\;=\;0\,, \tag{3.19}$$

where we have defined $\overline{a}(t,x;\delta u) \;:=\; \delta u_1 + x\circ\delta u_2$.

Proof. For any $0<|\varepsilon|<1$ small enough, let us define

$$\delta G^\varepsilon \;:=\; \frac{1}{\varepsilon}\Big(G(u+\varepsilon\delta u)\;-\;G(u)\Big).$$

It is easy to see that δG^ε solves the equation

$$\partial_t\delta G^\varepsilon\;+\;\operatorname{div}\Big(a(t,x;u)\,\delta G^\varepsilon\Big)\;=\;-\;\operatorname{div}\Big(\overline{a}(t,x;\delta u)\,G(u+\varepsilon\delta u)\Big)\,, \tag{3.20}$$

with initial datum $\delta G^\varepsilon_{|t=0}=0$.

Notice that, by uniform bounds provided by Lemma 3.1 (which holds for $m\geq 1$ and $k\geq 1$) and weak compactness methods, we can prove that δG^ε converges (up to extraction of a suitable subsequence) to some $f\in L^\infty_T(L^2)$ in the weak-$*$ topology of that space. Furthermore, this f satisfies the same equation as (3.20), with right-hand side equal to $-\operatorname{div}\Big(\overline{a}(t,x;\delta u)\,G(u)\Big)\in L^1_T(L^2)$. Now, by uniqueness we deduce that f has to coincide with $\delta_{\delta u}G$, and in addition the whole sequence $\Big(\delta G^\varepsilon\Big)_\varepsilon$ converges to it.

Unfortunately, the previous argument does not yield the Gâteaux differentiability of G, because we need that the limit exists in the strong topology, namely in the $L^\infty_T(L^2)$ norm. In order to get this property, let us write the equation for $f^\varepsilon\;:=\;\delta G^\varepsilon - f$, since $G(u+\varepsilon\delta u)-G(u)\;=\;\varepsilon\,\delta G^\varepsilon$, we find

$$\partial_t f^\varepsilon\;+\;\operatorname{div}\Big(a(x,t;u)\,f_\varepsilon\Big)\;=\;-\,\varepsilon\;\operatorname{div}\Big(\overline{a}(x,t;\delta u)\,\delta G^\varepsilon\Big)\,,$$

with zero initial datum. To formulate this equation, we need $m \geq 2$, $k \geq 2$. Then, an energy estimate immediately gives

$$\begin{aligned}\|f^\varepsilon(t)\|_{L^2} &\leq C\,\varepsilon\,\exp\left(C\int_0^t \|\operatorname{div} a(s,x;u)\|_{L^\infty}\right)\int_0^t \left\|\operatorname{div}\left(\overline{a}(s,x;\delta u)\,\delta G^\varepsilon\right)\right\|_{L^2} ds \\ &\leq C\,\varepsilon\,\exp\left(C\int_0^t \|\operatorname{div} a(s,x;u)\|_{L^\infty}\right)\int_0^t |\delta u(s)|\;\|\delta G^\varepsilon\|_{H^1_1}\,ds\,,\end{aligned}$$

where we have argued as in the first line of (3.13) in order to pass from the first inequality to the second one. At this point, applying the second estimate of Lemma 3.1 to equation (3.20) yields

$$\|\delta G^\varepsilon(s)\|_{H^1_1} \leq C_0\int_0^s |\delta u(s)|\,ds\,, \tag{3.21}$$

for any $s \in [0,t]$, $t \leq T$, for a fixed constant C_0 (depending on T, u^a, u^b, and $\|\nabla a_0\|_{L^1_T(C^2_b)}$, $\|f_0\|_{H^2_2}$ and $\|g\|_{L^1_T(H^2_2)}$). Putting this bound in the previous estimate entails

$$\begin{aligned}\|f^\varepsilon(t)\|_{L^2} &\leq C\,C_0\,\varepsilon\,\exp\left(C\int_0^t \|\operatorname{div} a(s,x;u)\|_{L^\infty}\right)\left(\int_0^t |\delta u(s)|\,ds\right)^2 \\ &\leq \varepsilon\,C\,\|\delta u\|^2_{L^\infty_T}\,\exp\left(C\|\operatorname{div} a(t,x;u)\|_{L^1_T(L^\infty)}\right)\,,\end{aligned}$$

from which we deduce that $f^\varepsilon \to 0$ in $L^\infty_T(L^2)$ for $\varepsilon \to 0$. The proposition is now proved. □

Next, we tackle the proof of the Fréchet differentiability of G.

Theorem 3.1. *Let $m \geq 2$ and $k \geq 2$. Let the data f_0, g and a_0 be fixed as in Remark 3.1 above, and let $u \in \operatorname{int} U_{ad}$. Define $DG(u)[\delta u]$ to be the unique solution to equation (3.19).*

Then there exists a constant $C > 0$ (depending only on T, u^a, u^b, and $\|\nabla a_0\|_{L^1_T(C^2_b)}$, $\|f_0\|_{H^2_2}$ and $\|g\|_{L^1_T(H^2_2)}$) such that

$$\left\|G(u+\delta u)\;-\;G(u)\;-\;DG(u)[\delta u]\right\|_{L^\infty_T(L^2)} \leq C\;\|\delta u\|^2_{L^\infty_T}\,.$$

In particular, the map G is Fréchet differentiable from $\operatorname{int} U_{ad}$ into $L^\infty_T(L^2)$, and its Fréchet differential at any point $u \in \operatorname{int} U_{ad}$ is given by $DG(u)$.

Proof. In order to prove that G is Fréchet differentiable, with Fréchet differential given by $DG(u)[\delta u]$, we have to show that

$$\lim_{\|\delta u\|_{L^\infty_T}\to 0}\frac{\left\|G(u+\delta u)-G(u)-DG(u)[\delta u]\right\|_{L^\infty_T(L^2)}}{\|\delta u\|_{L^\infty_T}} = 0\,.$$

We recall also that, if G is Fréchet differentiable at u, then it is also Gâteaux differentiable at the same point, and one has $\delta_{\delta u}G = DG(u)[\delta u]$.

For simplicity, let us introduce the notation $\mathcal{G}_u(\delta u) := G(u+\delta u) - G(u) - DG(u)[\delta u]$. The same computations performed on f^ε, in the proof of Proposition 3.3 above, lead us to the following equation for $\mathcal{G}_u(\delta u)$

$$\partial_t \mathcal{G}_u(\delta u) \,+\, \operatorname{div}\Big(a(t,x;u)\,\mathcal{G}_u(\delta u)\Big) \;=\; -\operatorname{div}\Big(\overline{a}(t,x;\delta u)\,\Big(G(u+\delta u) - G(u)\Big)\Big)\,,$$

with initial datum $\mathcal{G}_u(\delta u)_{|t=0} = 0$. Now, it is just a matter of repeating the estimates performed on f^ε: we easily find, for every $t \in [0,T]$, the inequality

$$\begin{aligned}\left\|\mathcal{G}_u(\delta u)(t)\right\|_{L^2} \;\leq\; C\,\exp\left(C\int_0^t \|\operatorname{div} a(s,x;u)\|_{L^\infty}\right)\,\times&\\ \times\int_0^t |\delta u(s)|\;\left\|G(u+\delta u) - G(u)\right\|_{H^1_1}\,ds\,.&\end{aligned}$$

Observe that an inequality analogous to (3.21) holds also for $G(u+\delta u) - G(u)$: inserting this relation in the previous estimate, we find

$$\begin{aligned}\left\|\mathcal{G}_u(\delta u)(t)\right\|_{L^2} \;&\leq\; C\,C_0\,\exp\left(C\int_0^t \|\operatorname{div} a(s,x;u)\|_{L^\infty}\right)\,\left(\int_0^t |\delta u(s)|\,ds\right)^2\\ &\leq\; K\,\|\delta u\|^2_{L^\infty_T}\,,\end{aligned}$$

for a new positive constant K. From this last inequality, the claims of the theorem follow. □

3.3. Analysis of Liouville ensemble optimal control problems

In this section, we investigate our Liouville ensemble optimal control problem. In the first part, we prove the existence of optimal controls by means of classical arguments. However, notice that one has to carefully justify that the reduced functional J_r (see its definition below) is weakly lower semi-continuous. In fact, this property is not obvious, since $f = G(u)$ depends non-linearly on u. After that, in Section 3.3.2, we characterize optimal controls as solutions of a related first-order optimality system. In Section 3.3.3 we discuss uniqueness of optimal controls under certain assumptions.

3.3.1. Existence of optimal controls

In this section, we deal with existence of optimal solutions to an ensemble optimal control problem. Our analysis is based on the following assumptions.

(A.1) We fix $(m,k) \in \mathbb{N}^2$, and we take an initial datum $f_0 \in H^m_k(\mathbb{R}^d)$.

(A.2) We fix parameters $(\gamma,\delta,\nu) \in \mathbb{R}^3$ such that $\gamma > 0$, $\delta \geq 0$ and $\nu \geq 0$.

(A.3) Chosen $u^a = \left(u_1^a, u_2^a\right)$ and $u^b = \left(u_1^b, u_2^b\right)$ in $\mathbb{R}^{2d}$, with $u^a \leq u^b$, we define the set of admissible controls to be

$$U_{ad} := \left\{ u \in \mathbb{L}_T^\infty(\mathbb{R}^d) \;\middle|\; u^a \leq u(t) \leq u^b \quad \text{for a.e. } t \in [0,T] \right\} \quad \text{if} \quad \nu = 0 \tag{3.22}$$

$$U_{ad} := \left\{ u \in \mathbb{H}_T^1(\mathbb{R}^d) \;\middle|\; u^a \leq u(t) \leq u^b \quad \text{for all } t \in [0,T] \right\} \quad \text{if} \quad \nu > 0\,. \tag{3.23}$$

(A.4) We take two attracting potentials θ and φ in $L^2(\mathbb{R}^d)$, in the sense specified in Section 3.1.

Remark 3.3. *We point out that assumption **(A.4)** (which will be strengthened in Section 3.3.3 for getting uniqueness, see condition **(A.4)*** there) is taken for simplicity of presentation, since more general θ and φ can be considered in our framework. For instance, we can allow for θ to depend on time: $\theta \in L_T^1(L^2)$, or $\theta \in L_T^1(H_1^1)$ in **(A.4)*** below. The case $\theta(x) = |x|^2$ and $\varphi(x) = |x|^2$ is more delicate, and will be matter of further discussions in Section 3.3.4.*

Now, consider our cost functional given by

$$\begin{aligned} J(f,u) := &\int_0^T\!\!\int_{\mathbb{R}^d} \theta(x)\, f(x,t)\, dx\, dt + \int_{\mathbb{R}^d} \varphi(x)\, f(x,T)\, dx \\ &+ \frac{\gamma}{2} \int_0^T \left|u(t)\right|^2 dt + \delta \int_0^T \left|u(t)\right| dt + \frac{\nu}{2} \int_0^T \left|\frac{\mathrm{d}}{\mathrm{d}t} u(t)\right|^2 dt\,. \end{aligned} \tag{3.24}$$

We remark that J is well-defined whenever $u \in \mathbb{L}_T^2$ if $\nu = 0$, or $u \in \mathbb{H}_T^1$ if $\nu > 0$, and $f \in C\Big([0,T]; L^2(\mathbb{R}^d)\Big)$.

Our ensemble optimal control problem requires to find

$$\min_{u \in U_{ad}} J(f,u)\,, \tag{3.25}$$

subject to the differential constraint

$$\begin{cases} \partial_t f + \operatorname{div}\Big(a(x,t;u)\, f\Big) = g & \text{in} \quad \mathbb{R}^d \times (0,T] \\ f_{|t=0} = f_0 & \text{on} \quad \mathbb{R}^d\,, \end{cases} \tag{3.26}$$

where the drift function $a(x,t;u)$ is defined as

$$a(x,t;u) := a_0(x,t) + u_1(t) + x \circ u_2(t)\,. \tag{3.27}$$

That is taking $a_1 = a_2 = 1$ in (2.5).

Under our assumptions, Theorem 2.4 applies. Thus, for every $u \in U_{ad}$, there exists a unique corresponding solution $f \in C\Big([0,T]; H_k^m(\mathbb{R}^d)\Big)$ to the problem (3.26). Therefore, resorting to the control-to-state map G defined in Section 3.2, we can introduce the so-called reduced cost functional, given by

$$J_r(u) := J\Big(G(u), u\Big)\,. \tag{3.28}$$

Hence, the ensemble optimal control problem (3.25)–(3.26) can be rephrased as follows

$$\min_{u\in U_{ad}} J_r(u)\,. \tag{3.29}$$

Remark 3.4. *Recall that we have defined G with values in $L^\infty_T(L^2)$. However, under our assumptions, we know that the solution to the Liouville equation actually belongs to $C_T(L^2)$, so that the φ-term in* (3.24) *is well-defined, and so is J_r.*

In the following, we prove existence of a minimizer to (3.29).

Theorem 3.2. *Under assumptions **(A.1)-(A.2)-(A.3)-(A.4)**, the ensemble optimal control problem* (3.29) *admits at least one solution $u^* \in U_{ad}$. The corresponding state $f^* := G(u^*)$ belongs to the space $C\Big([0,T]; H^m_k(\mathbb{R}^d)\Big)$.*

Proof. Let us focus on the case $\nu = 0$ for simplicity; the case $\nu > 0$ follows from the same argument.

The functional J given in (3.24) is well-defined for $(f, u) \in C_T(L^2) \times \mathbb{L}^\infty_T$, and U_{ad} is a bounded subset of $\mathbb{L}^\infty_T$. On the other hand, owing to estimate (2.24) in Theorem 2.4, and the embedding $C_T(H^m_k) \hookrightarrow L^\infty_T(L^2)$, the map G takes its values in a bounded set of $L^\infty_T(L^2)$. It follows that J_r is bounded; in particular, J_r is a proper map, i.e., $\inf_{U_{ad}} J_r > -\infty$, and J_r is not identically equal to $+\infty$.

Next, we claim that J_r is weakly lower semi-continuous. To prove this fact, it is enough to use the weak-weak continuity of G, as stated in Proposition 3.2, and to remark that J is weakly lower semi-continuous. Indeed, the last three terms in (3.24) are norms, so they are weakly lower semi-continuous. On the other hand, the first two terms are linear in f, and then they are weakly continuous with respect to the $L^\infty_T(L^2)$ and L^2 topologies, respectively. Thus, we immediately get that, if $\Big(u_n\Big)_n \subset U_{ad}$ is a sequence which converges weakly-$*$ to a $u \in U_{ad}$ in L^∞_T, we have

$$\liminf_{n\to+\infty} J_r(u_n) \;=\; \liminf_{n\to+\infty} J\Big(G(u_n), u_n\Big) \;\geq\; J\Big(G(u), u\Big) \;=\; J_r(u)\,.$$

At this point, proving the existence of a minimizer for J_r is standard. Let us take a minimizing sequence $\Big(u_n\Big)_n \subset U_{ad}$. Since U_{ad} is a bounded set in $\mathbb{L}^\infty_T$, we can extract a weakly-$*$ convergent subsequence, which we do not relabel for simplicity. Let us call $u^* \in U_{ad}$ its limit-point. Then, by the weak-lower semi-continuity of J_r, we can conclude that u^* is a minimizer for J_r. □

We discuss uniqueness of the minimizers in Section 3.3.3 below. For this purpose, we use characterization of minimizers as solutions to a suitable optimality system, which we derive in the next section.

3.3.2. Liouville optimality systems

This section is devoted to the characterization of ensemble optimal controls as solutions of the related first-order optimality system. For this purpose, in addition to hypotheses **(A.1)**–**(A.2)**–**(A.3)**–**(A.4)** stated above, from now on we take

$$m \geq 1 \qquad \text{and} \qquad k \geq 1\,.$$

In correspondence to (3.24)–(3.25)–(3.26), we consider the Lagrange multipliers framework, see e.g. [**94, 126**], and introduce the Lagrange functional $\mathcal{L}$ as follows:

(3.30)

$$\begin{aligned}\mathcal{L}(f,u,q) \;:=\; J(f,u) &+ \int_0^T\!\!\int_{\mathbb{R}^d} \Big(\partial_t f(x,t) + \operatorname{div}\big(a(x,t;u)f(x,t)\big) - g(x,t)\Big)q(x,t)\,dx\,dt \\ &+ \int_{\mathbb{R}^d} \big(f(0,x) - f_0(x)\big)q_0(x)\,dx\,,\end{aligned}$$

where, for the sake of generality, we have included a right-hand side g. The variable q represents the Lagrange multiplier. Notice that $\mathcal{L}$ is well-defined whenever $u \in \mathbb{L}^\infty_T$ if $\nu = 0$, $u \in \mathbb{H}^1_T$ if $\nu > 0$, $q \in L^\infty_T(L^2)$, $q_0 \in L^2$ and $f \in C_T(L^2)$ such that both $\partial_t f$ and $\operatorname{div}\big(a(x,t;u)\,f\big)$ belong to $L^1_T(L^2)$. In particular, it is enough to have $f \in W^{1,1}_T(L^2) \cap L^\infty_T(H^1_1)$, recall also Proposition 2.2. Notice that, *a posteriori*, we will find $q \in C_T(L^2)$ and $q_0 = q(0)$; see the discussion below for details.

For clarity, in order to derive the optimality system, we first discuss the case with L^2 costs only, then the case with $L^2 - H^1$ costs, and finally the case with $L^2 - L^1 - H^1$ costs.

The case $\boldsymbol{\delta = \nu = 0}$. If $\delta = 0$, then J is Fréchet differentiable over $C_T(L^2) \times \operatorname{int} U_{ad}$, since it is linear in f and the control costs with $\gamma > 0$, $\nu \geq 0$ are given by differentiable norms. It follows then that $\mathcal{L}$ is Fréchet differentiable over the space

$$\mathbb{X}_T \;:=\; \Big(W^{1,1}_T(L^2) \cap L^\infty_T(H^1_1)\Big) \;\times\; \mathbb{L}^2_T \;\times\; C_T(L^2)\,,$$

where $\mathbb{L}^2_T$ has to be replaced by $\mathbb{H}^1_T$ in the case when $\nu > 0$. The Fréchet differential of $\mathcal{L}$ at (f,u,q) is given by the linearization of each of its terms at that point.

Now, consider in addition $\nu = 0$. The optimality system is obtained by putting to zero the Fréchet derivatives of $\mathcal{L}(f,u,q)$ with respect to each of its arguments separately. We obtain

$$\text{(3.31)}\quad \partial_t f \,+\, \operatorname{div}\Big(a(x,t;u)\,f\Big) \;=\; g\,, \qquad \text{with} \quad f_{|t=0} \;=\; f_0$$

$$\text{(3.32)}\quad -\,\partial_t q \,-\, a(x,t;u)\cdot\nabla q \;=\; -\,\theta, \qquad \text{with} \quad q_{|t=T} \;=\; -\,\varphi$$

$$\text{(3.33)}\quad \left(\gamma\, u^j_\iota + \int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u^j_\iota}\, f\right) q\,dx,\; v^j_\iota - u^j_\iota\right)_{L^2(0,T)} \geq 0 \quad \forall v \in U_{ad}, \iota = 1,2,\; j = 1\ldots d\,.$$

We remark that, denoting by e^j the j-th unit vector of the canonical basis of $\mathbb{R}^d$ and by x^j the j-th component of the vector $x \in \mathbb{R}^d$, we have $\partial a / \partial u_1^j = e^j$ and $\partial a / \partial u_2^j = x^j \, e^j$ by Definition 3.27. Then, equation (3.33) can be equivalently written in the following form. For any $1 \leq j \leq d$, we have

$$\begin{cases} \Big(\gamma \, u_1^j + \displaystyle\int_{\mathbb{R}^d} \partial_j f \, q \, dx \; , \; v_1^j - u_1^j \Big)_{L^2(0,T)} \geq 0 \\ \Big(\gamma \, u_2^j + \displaystyle\int_{\mathbb{R}^d} \partial_j \Big(x^j \, f \Big) \, q \, dx \; , \; v_2^j - u_2^j \Big)_{L^2(0,T)} \geq 0 \, . \end{cases}$$

Further, if we sum up equations (3.33) for all ι and all j, we can write

$$\Big(\gamma \, u + \int_{\mathbb{R}^d} \operatorname{div} \Big((e + x) \, f \Big) \, q \, dx \; , \; v - u \Big)_{\mathbb{L}^2_T} \geq 0 \qquad \text{for all} \quad v \in U_{ad} \, , \tag{3.34}$$

where we have defined the vector $e = (1, \dots, 1)^T$.

Equation (3.31) is our Liouville model and is also called the forward equation in this context. The results of Section 2.2.2 guarantee that, under our assumptions, there exists a unique solution $f \in C_T(H^1_1)$. Moreover, since $u \in U_{ad}$, an inspection of (3.31) reveals that $\partial_t f \in L^1_T(L^2)$.

Equation (3.32) is the adjoint Liouville equation. It is obtained by taking the Fréchet derivative of (3.30) with respect to f. This is a transport equation that evolves backwards in time. By setting $\widetilde{q}(t, x) = q(T - t, -x)$, we obtain a transport problem for $\widetilde{q}$, as in (2.21), with source term $-\theta$ and initial condition $\widetilde{q}_{|t=0} = -\varphi$. Thus, the results in Section 2.2 guarantee the existence and uniqueness of a Lagrange multiplier $q \in C_T(L^2)$, provided that θ and φ are in L^2.

From the discussion above, we get that any solution to the optimality system (3.31)–(3.32)–(3.33), with $u \in U_{ad}$, belongs indeed to the space $\mathbb{X}_T$.

Equation (3.33) represents the optimality condition. To better illustrate this fact, we suppose from now on that

$$m \geq 2 \qquad \text{and} \qquad k \geq 2 \, .$$

Then, the reduced cost functional J_r, defined in (3.28), is Fréchet differentiable. In terms of the reduced minimization problem (3.29), the optimal solution u^* in the convex, closed and bounded set U_{ad} is characterized by the optimality condition given by

$$(\nabla_u J_r(u^*) \; , \; v - u^*)_{\mathbb{L}^2_T} \geq 0, \qquad \text{for all } v \in U_{ad} \, ,$$

where $\nabla_u J_r$ denotes the L^2-gradient of J_r with respect to u. In fact, a direct computation of $\nabla_u J\Big(G(u), u\Big)$, with the introduction of the auxiliary adjoint variable q, gives the optimality system above, and the following relation

$$\nabla_{u_\iota^j} J_r(u) \;=\; \gamma\, u_\iota^j \;+\; \int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u_\iota^j}\, f \right) q\, dx \,.$$

The case $\boldsymbol{\delta = 0}$, $\boldsymbol{\nu > 0}$. Next, assume that $\delta = 0$ and $\gamma, \nu > 0$. Recall that, in this case, the set U_{ad} is defined by (3.23). Then, the natural Hilbert space where u^* is sought is $\widetilde{\mathbb{H}}_T^1(\mathbb{R}^d) \;:=\; \widetilde{H}_T^1(\mathbb{R}^d) \times \widetilde{H}_T^1(\mathbb{R}^d)$, where $\widetilde{H}_T^1$ corresponds to the H_T^1 space, endowed with the weighted H^1-product given by

$$(u, v)_{\widetilde{H}_T^1} \;:=\; \gamma \int_0^T u(t) \cdot v(t)\, dt \;+\; \nu \int_0^T u'(t) \cdot v'(t)\, dt \,.$$

The notation $\;' \;=\; d/dt$ stands for the weak time derivative.

Now, let μ be the $\widetilde{H}^1$-Riesz representative of the continuous linear functional

$$v \;\mapsto\; \left(\int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u}\, f \right) q\, dx \;,\; v \right)_{\mathbb{L}_T^2} \,.$$

Assuming that $u \;\in\; U_{ad} \cap H_0^1\Big([0,T]; \mathbb{R}^{2d}\Big)$, then μ can be computed by solving the boundary-value problem

$$\left(-\, \nu\, \frac{d^2}{dt^2} \;+\; \gamma \right) \mu \;=\; \int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u}\, f \right) q\, dx \,, \qquad \mu(0) \;=\; \mu(T) \;=\; 0 \,, \tag{3.35}$$

which is understood in a weak sense. Notice that the choice $u \;\in\; H_0^1\Big([0,T]; \mathbb{R}^{2d}\Big)$ corresponds to the modelling requirement that the control is switched on at $t \;=\; 0$ and switched off at $t \;=\; T$. Other initial and final time conditions on u may be required and encoded as boundary conditions in (3.35).

With the setting above, the $\widetilde{H}^1$-gradient is given, for $\iota = 1, 2$ and $j = 1 \ldots d$, by

$$\widetilde{\nabla}_{u_\iota^j} J_r(u) \;=\; u_\iota^j \;+\; \mu_\iota^j \,. \tag{3.36}$$

The optimality condition (3.33) then becomes

$$\Big(u_\iota^j \;+\; \mu_\iota^j \;,\; v_\iota^j \;-\; u_\iota^j \Big)_{\widetilde{H}_T^1} \;\geq\; 0, \tag{3.37}$$

for all $v \;\in\; U_{ad}$, $\iota = 1, 2$ and $1 \leq j \leq d$.

The case $\boldsymbol{\delta > 0}$. In this case, a L^1 norm of the control appears in the cost functional. This term is not Gâteaux differentiable and the discussion becomes more involved. Using the control-to-state map, we start by defining

$$\begin{aligned} j_1(u) \;:=\; & \int_0^T\!\!\int_{\mathbb{R}^d} \theta(x)\, G(u)(x,t)\, dx\, dt \;+\; \int_{\mathbb{R}^d} \varphi(x)\, G(u)(x,T)\, dx \\ & \qquad\qquad +\; \frac{\gamma}{2} \int_0^T \Big| u(t) \Big|^2 dt \;+\; \frac{\nu}{2} \int_0^T \left| \frac{d}{dt} u(t) \right|^2 dt, \\ j_2(u) \;:=\; & \delta\, \|u\|_{L_T^1} \,. \end{aligned}$$

The L^1-cost, represented by j_2, admits a subdifferential $\partial j_2(u) = \delta\,\partial\Big(\|u\|_{L^1} \Big)$; see, e.g., Section 2.3 of [**11**]. If we denote by $\mathbb{L}_T^* := \Big(\mathbb{L}_T^\infty(\mathbb{R}^d)\Big)^*$ and by $\langle\cdot,\cdot\rangle$ the duality product in $\mathbb{L}_T^* \times \mathbb{L}_T^\infty$, the following formula holds true:

$$\begin{aligned} \partial\Big(\|u\|_{L^1} \Big) &= \Big\{ \phi \in \mathbb{L}_T^* \;\Big|\; \|v\|_{L^1} - \|u\|_{L^1} \geq \Big\langle \phi\,,\, v-u \Big\rangle \quad \forall\, v \in U_{ad} \Big\} \\ &= \begin{cases} \Big\{ \phi \in \mathbb{L}_T^* \;\Big|\; \|\phi\|_{\mathbb{L}_T^*} = 1\,,\; \langle\phi,u\rangle = \|u\|_{\mathbb{L}_T^\infty} \Big\} & \text{if} \quad u \not\equiv 0 \\ \text{unit ball in } \mathbb{L}_T^* & \text{if} \quad u \equiv 0\,. \end{cases} \end{aligned} \tag{3.38}$$

Now, the reduced functional can be written as $J_r(u) = j_1(u) + j_2(u)$. In this case, the equations (3.31) and (3.32) in the corresponding optimality system are the same. However, we have a different optimality condition (3.33). In the case $\nu = 0$, as in Theorem 2.2 in [**49**], we have the following Theorem 3.3. For its proof, we refer to [**49**] and [**119**]. Notice that, as for equations (3.31)–(3.32)–(3.33), equation (3.39) below can be written even when G, and hence J_r, are not Fréchet differentiable.

Theorem 3.3. *Under assumptions* ***(A.1)–(A.2)–(A.3)–(A.4)****, where we take $m \geq 1$ and $k \geq 1$, we suppose that the pair $(f,u) \in C_T(H_k^m) \times U_{ad}$ is a minimizer for* (3.29)*.*

Then there exists a unique $q \in C_T(L^2)$ which solves (3.32)*, and a $\widehat{\lambda} \in \partial g(u)$ such that the following inequality condition is satisfied:*

$$\left(\gamma\, u_\iota^j + \widehat{\lambda}_\iota^j + \int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u_\iota^j}\, f \right) q\, dx, v_\iota^j - u_\iota^j \right)_{L^2(0,T)} \geq 0 \quad \forall\, v \in U_{ad}\,,\; \iota = 1,2, j = 1\ldots d\,. \tag{3.39}$$

Moreover, there exist λ_+ and λ_-, belonging to $L_T^\infty(\mathbb{R}^d)$, such that (3.39) *is equivalent to the equations*

$$\begin{cases} \gamma\, u_\iota^j + \displaystyle\int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u_\iota^j}\, f \right) q\, dx + (\lambda_+)_\iota^j - (\lambda_-)_\iota^j + \widehat{\lambda}_\iota^j = 0 \\ (\lambda_+)_\iota^j \geq 0\,, \qquad u^b - u_\iota^j \geq 0\,, \qquad (\lambda_+)_\iota^j\,(u^b - u_\iota^j) = 0 \\ (\lambda_-)_\iota^j \geq 0\,, \qquad u_\iota^j - u^a \geq 0\,, \qquad (\lambda_-)_\iota^j\,(u_\iota^j - u^a) = 0 \\ \widehat{\lambda}_\iota^j = \delta \qquad a.e.\ in \quad \Big\{ t \in [0,T] \;\Big|\; u_\iota^j(t) > 0 \Big\} \\ \Big|\widehat{\lambda}_\iota^j\Big| \leq \delta \qquad a.e.\ in \quad \Big\{ t \in [0,T] \;\Big|\; u_\iota^j(t) = 0 \Big\} \\ \widehat{\lambda}_\iota^j = \delta \qquad a.e.\ in \quad \Big\{ t \in [0,T] \;\Big|\; u_\iota^j(t) < 0 \Big\}\,, \end{cases} \tag{3.40}$$

for $\iota = 1,2$ and all $1 \leq j \leq d$.

In (3.40), one usually refers to the first equation as the optimality condition equation; the conditions given in the second and third line are the complementarity conditions for the inequality constraints in U_{ad}. Moreover, the last three lines give an equivalent expression for $\widehat{\lambda} \in \partial g(u)$; see [**119**]. In our case, $\widehat{\lambda}_m^r$ can be understood to be $\delta \operatorname{sgn}(u_m^r)$, where $\operatorname{sgn}(x)$ is the sign function.

Finally, **the case $\boldsymbol{\delta} > 0$ and $\boldsymbol{\nu} > 0$** can be treated as done before. After resorting once again to the space $\widetilde{\mathbb{H}}_T^1$, let μ be the $\widetilde{H}^1$-Riesz representative of the continuous linear functional

$$v \;\mapsto\; \left(\widehat{\lambda} + \int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u} f\right) q \, dx \;,\; v\right)_{\mathbb{L}_T^2} .$$

Then, assuming that $u \in U_{ad} \cap H_0^1\Big([0,T]; \mathbb{R}^{2d}\Big)$, we can compute μ as above, by solving the boundary-value problem

$$\left(-\nu \frac{d^2}{dt^2} + \gamma\right)\mu \;=\; \widehat{\lambda} + \int_{\mathbb{R}^d} \operatorname{div}\left(\frac{\partial a}{\partial u} f\right) q \, dx \,, \qquad \mu(0) \;=\; \mu(T) \;=\; 0\,. \tag{3.41}$$

With this definition, relation (3.36) still holds true, and the optimality condition (3.33) can be expressed once again by equations (3.37).

3.3.3. Uniqueness of optimal controls

In this section, we prove uniqueness of optimal controls in the situation when $\delta = 0$ and $\nu = 0$ in (3.24). Our proof relies on the characterization of optimal controls as solutions to the corresponding optimality system. The cases $\delta > 0$ or $\nu > 0$ read more complicated and are left aside in our discussion.

To begin with, we need additional regularity on the cost functions θ and φ in order to prove uniqueness. We formulate the following assumption, which strengthens **(A.4)**:

(A.4)* Suppose that both θ and φ belong to $H_1^1(\mathbb{R}^d)$.

In the constrained-control case, the characterization of optimal controls is given by an inequality, see (3.33). This is a very weak information. This is the reason why we are able to prove uniqueness only under a smallness condition, either on the time T or on the size of the data f_0, g, ∇a_0, θ and φ in their respective functional spaces. Let us recall that existence of an optimal control has been proved in Theorem 3.2 above.

Theorem 3.4. *Under assumptions **(A.1)**-**(A.2)**-**(A.3)**-**(A.4)***, suppose that both $m \geq 2$ and $k \geq 2$. Take moreover $\delta = \nu = 0$ in* (3.24). *Finally, define*

$$\widetilde{K} \;:=\; C \,\exp\Big(C\left(\|\nabla a_0\|_{L_T^1(C_b^2)} + T\,\max\left\{\left|u^a\right|,\,\left|u^b\right|\right\}\right)\Big) \\ \times\left(\|f_0\|_{H_2^2} + \|g\|_{L_T^1(H_2^2)}\right)\left(\|\varphi\|_{H_1^1} + T\,\|\theta\|_{H_1^1}\right),$$

where the constant $C > 0$ *can be taken as the maximum of the constants* C *appearing in* (3.43)*,* (3.45)*,* (3.46) *and in the definition* (3.11) *of* $K_1^{(2)}$*.*

If the condition $\widetilde{K}\,T/\gamma\,<\,1$ *holds true, then there exists at most one optimal control* u^* *in* $\operatorname{int} U_{ad}$*.*

Proof. The previous result being classical in optimal control problems, let us just give a sketch of the proof. Let (u, f_1, q_1) and (v, f_2, q_2) be two optimal triplets solving the minimization problem (3.29). From (3.34) we deduce that, for all $w \in U_{ad}$,

$$\Big(\gamma u + \int_{\mathbb{R}^d} \operatorname{div}\Big((e+x)f_1\Big)q_1\,,\; u - w\Big)_{\mathbb{L}^2_T} \leq 0$$

and

$$\Big(\gamma v + \int_{\mathbb{R}^d} \operatorname{div}\Big((e+x)f_2\Big)q_2\,,\; w - v\Big)_{\mathbb{L}^2_T} \geq 0\,.$$

Take $w = v$ in the former inequality, $w = u$ in the latter and compute the difference of the resulting expressions. After setting $\delta f \,:=\, f_1 - f_2$ and $\delta q \,:=\, q_1 - q_2$, straightforward computations lead to

$$\begin{aligned}(3.42)\quad \gamma\int_0^T |u(t)-v(t)|^2\,dt \,\leq\, \int_0^T\Big[\int_{\mathbb{R}^d}\Big|\operatorname{div}\Big((e+x)\,\delta f\Big)q_1\Big| \\ +\int_{\mathbb{R}^d}\Big|\operatorname{div}\Big((e+x)\,f_2\Big)\,\delta q\Big|\Big]\,|u(t)-v(t)|\,dt\,.\end{aligned}$$

Now we estimate the two space integrals, at any time $t \in [0,T]$. We start with the former term, for which we obtain

$$\begin{aligned}\int_{\mathbb{R}^d}\Big|\operatorname{div}\Big((e+x)\,\delta f(t)\Big)\,q_1(t)\Big|\,dx \,&\leq\, \|q_1(t)\|_{L^2}\,\Big\|\operatorname{div}\Big((e+x)\,\delta f(t)\Big)\Big\|_{L^2} \\ &\leq\, C_1\,\Big(\|\delta f(t)\|_{L^2}\,+\,\Big\|\Big(1+|x|\Big)\,\nabla\delta f(t)\Big\|_{L^2}\Big) \\ &\leq\, C_1\,\|\delta f(t)\|_{H^1_1}\,,\end{aligned}$$

where we have also used Theorem 2.3 applied to the transport equation (3.32) for treating the q_1 term. Notice that the constant C_1 can be expressed as

$$(3.43)\qquad C_1 \,:=\, C\,\exp\Big(C\,\Big(\|\operatorname{div} a_0\|_{L^1_T(L^\infty)} + \|u_1\|_{\mathbb{L}^1_T}\Big)\Big)\Big(\,\|\varphi\|_{L^2}\,+\,T\,\|\theta\|_{L^2}\,\Big)\,,$$

for a "universal" constant $C > 0$ that depends on the space dimension d. At this point, we recall that both f_1 and f_2 satisfy equation (3.31), with controls u_1 and u_2, respectively. Then, taking their difference and applying Lemma 3.1 finally yields, for a new constant $\widetilde{C}_1 \,=\, C_1\,K_1^{(2)}$ just depending on the data of the problem, the following bound:

$$(3.44)\qquad \int_{\mathbb{R}^d}\Big|\operatorname{div}\Big((e+x)\,\delta f(t)\Big)q_1(t)\Big|\,dx \,\leq\, \widetilde{C}_1\int_0^t\Big|u(s)-v(s)\Big|\,ds\,.$$

Next, consider the second integral in (3.42). The computations are similar to the previous ones. We can estimate

$$\int_{\mathbb{R}^d} \Big|\mathrm{div}\,\Big((e+x)f_2(t)\Big)\,\delta q(t)\Big|\,dx \,\leq\, \|\delta q(t)\|_{L^2}\,\|f_2(t)\|_{H^1_1} \,\leq\, C_2\,\|\delta q(t)\|_{L^2}\,,$$

where we have applied Theorem 2.4 to equation (3.31) for f_2 to control its H^1_1 norm. In particular, it follows from that theorem that

$$C_2 \,:=\, C\,\exp\Big(C\,\Big(\|\nabla a_0\|_{L^1_T(C^1_b)} \,+\, \|u_2\|_{\mathbb{L}^1_T}\Big)\Big)\,\Big(\|f_0\|_{H^1_1} \,+\, \|g\|_{L^1_T(H^1_1)}\Big)\,, \tag{3.45}$$

for a "universal" constant $C > 0$.

Now, we use the fact that q_1 and q_2 are both solutions of (3.32), related to the controls u_1 and u_2 respectively. Hence, taking the difference of the corresponding equations and arguing as in the proof of Lemma 3.1, one easily infers the existence of a "universal" constant $C > 0$ such that

$$\begin{aligned}\|\delta q(t)\|_{L^2} \,\leq\, & C\,\exp\Bigg(C\,\Big(\|\mathrm{div}\,a_0\|_{L^1_T(L^\infty)} \,+\, \|u_1\|_{\mathbb{L}^1_T}\Big)\Bigg)\times \\ & \qquad\qquad \times\int_t^T |u(s)-v(s)|\,\Big\|\Big(1+|x|\Big)\,\nabla q_2(s)\Big\|_{L^2}\,ds \\ \leq\, & C\,\exp\Bigg(C\,\Big(\|\nabla a_0\|_{L^1_T(C^1_b)} + \|u_1\|_{\mathbb{L}^1_T} + \|u_2\|_{\mathbb{L}^1_T}\Big)\Bigg)\times \\ & \qquad\qquad \times\Big(\|\varphi\|_{H^1_1} + T\,\|\theta\|_{H^1_1}\Big)\int_t^T \Big|u(s)-v(s)\Big|\,ds\,.\end{aligned}$$

Notice that the integral is from t to T, because (3.32) is a backward transport equation. After defining the constants

$$\widetilde{K}^{(1)}_1 \,:=\, C\,\exp\Bigg(C\,\Big(\|\nabla a_0\|_{L^1_T(C^1_b)} + \|u_1\|_{\mathbb{L}^1_T} + \|u_2\|_{\mathbb{L}^1_T}\Big)\Bigg)\,\Big(\|\varphi\|_{H^1_1} + T\,\|\theta\|_{H^1_1}\Big) \tag{3.46}$$

and $\widetilde{C}_2 \,:=\, C_2\,\widetilde{K}^{(1)}_1$, we obtain

$$\int_{\mathbb{R}^d} \Big|\mathrm{div}\,\Big((e+x)\,f_2(t)\Big)\,\delta q(t)\Big|\,dx \,\leq\, \widetilde{C}_2\,\int_t^T \Big|u(s)-v(s)\Big|\,ds\,. \tag{3.47}$$

At this point, we can insert estimates (3.44) and (3.47) into (3.42), and get, for a new constant $K \,=\, \widetilde{C}_1 + \widetilde{C}_2$, the relation

$$\gamma\int_0^T \Big(\sigma(t)\Big)^2\,dt \,\leq\, K\,\int_0^T \sigma(t)\,\left(\int_0^T \sigma(t)\,ds\right)\,dt \,=\, K\left(\int_0^T \sigma(t)\,dt\right)^2\,,$$

where, for simplicity of notation, we have defined $\sigma(t) \,:=\, \Big|u(t)-v(t)\Big|$. Hence, by Cauchy-Schwarz inequality we easily deduce

$$\gamma\int_0^T \Big(\sigma(t)\Big)^2\,dt \,\leq\, K\,T\int_0^T \Big(\sigma(t)\Big)^2\,dt\,,$$

which obviously implies $\sigma \equiv 0$ almost everywhere on $[0,T]$ whenever $K\,T/\gamma \,<\, 1$. Then, we conclude the proof remarking that $K \,\leq\, \widetilde{K}$. □

3.3.4. The case of confining θ and φ as quadratic functions

As pointed out in Remark 3.3, from the applications viewpoint, it may be desirable to consider the case when both θ and φ are quadratic potentials. In this section, we discuss the necessary adaptations to be implemented in our arguments in order to address this case.

Therefore, from now on we choose

$$\theta(x) \,=\, |x|^2 \qquad \text{and} \qquad \varphi(x) \,=\, |x|^2 \,,$$

although the discussion can be further adapted, in order to treat more general polynomial growths. In order to simplify the presentation, we also assume that $\delta = \nu = 0$.

We notice that, in view of (3.24), for J to be well-defined it is necessary that $|x|^2\,f$ belongs to L^1. Then, we have to assume higher integrability on f, namely that

$$f \,\in\, C\Big([0,T];L^2_k(\mathbb{R}^d)\Big)\,, \qquad\qquad \text{for some} \quad k \,>\, 2+\frac{d}{2}\,.$$

This of course entails that, in **(A.1)**, one has to take $f_0 \,\in\, H^m_k$ and $g \,\in\, L^1_T(H^m_k)$, with the same restriction $k > 2 + d/2$. However, Theorem 3.2 still holds true.

The main changes pertain Section 3.3.2, starting from the Definition 3.30 of the functional $\mathcal{L}$. To begin with, let us focus on the Lagrangian multiplier q. On the one hand, we need it to be in some duality pairing with f. Then, keeping in mind Definition 2.1, we introduce, for $(m,k) \in \mathbb{N}^2$, the spaces

$$H^m_{-k}(\mathbb{R}^d) \,:=\, \Big\{f \in H^m_{\rm loc}(\mathbb{R}^d) \;\Big|\; \Big(1+|x|\Big)^{-k} D^\alpha f \,\in\, L^2(\mathbb{R}^d) \quad \forall\, 0 \leq |\alpha| \leq m\Big\}\,.$$

This space is endowed with the natural norm

$$\|f\|_{H^m_{-k}} \,=\, \sum_{0\leq|\alpha|\leq m} \left\|\Big(1+|x|\Big)^{-k} D^\alpha f\right\|_{L^2}\,.$$

On the other hand, we still expect q to solve (3.32) to an extent, although the meaning of that equation is now no more clear, owing to the fact that θ and φ do not belong anymore to L^2. To deal with both issues, we need the following lemma, whose proof can be performed arguing as in the proof of Theorem 2.4 above, using this time the weight $\Big(1+|x|\Big)^{-k}$. We omit to give the details here.

Lemma 3.2. *Let $T > 0$ and $(m,k) \in \mathbb{N}^2$ fixed, and let a be a vector field satisfying hypotheses* (2.11). *Moreover, assume that $q_0 \in H^m_{-k}(\mathbb{R}^d)$ and $g \in L^1\Big([0,T];H^m_{-k}(\mathbb{R}^d)\Big)$. Then there exists a unique solution $q \,\in\, C\Big([0,T];H^m_{-k}(\mathbb{R}^d)\Big)$ to the problem*

$$\partial_t q \,+\, a\cdot\nabla q \,=\, g\,, \qquad\qquad \textit{with} \quad q_{|t=0} \,=\, q_0\,.$$

Moreover, there exists a constant $C > 0$ such that the following estimate holds true for any $t \in [0,T]$:

$$\text{(3.48)} \quad \|q(t)\|_{H^m_{-k}} \leq C \exp\Big(C \int_0^t \|\nabla a(s)\|_{C^m_b} \, ds \Big) \Big(\|q_0\|_{H^m_{-k}} + \int_0^t \|g(s)\|_{H^m_{-k}} \, ds \Big) .$$

Let us come back to our optimal control problem. In view of Lemma 3.2, we can solve equation (3.32) with θ and φ equal to $|x|^2$, getting a unique solution in the space $C_T(L^2_{-k})$ for any $k > 2 + d/2$. Let us fix the choice

$$k_0 \, = \, 3 \, + \, \left[\frac{d}{2} \right] ,$$

where given $z \in \mathbb{R}$, we denote by $[z]$ its entire part. Then, it is easy to see that the functional $\mathcal{L}$ is well-defined on the space

$$\widetilde{\mathbb{X}}_T \, := \, \Big(W^{1,1}_T(L^2_{k_0}) \cap L^\infty_T(H^1_{k_0+1}) \Big) \, \times \, \mathbb{L}^2_T \, \times \, C_T(L^2_{-k_0}) .$$

Of course, we also need to take f_0 and g as in assumption **(A.1)**, with $m \geq 1$ and $k \geq k_0 + 1$.

Thereafter, we can write the optimality system (3.31)–(3.32)–(3.33), as done above. In order to characterize equation (3.33) in terms of the gradient of the reduced functional J_r, we need to further assume that $m \geq 2$ and $k \geq k_0 + 2$.

Finally, also the analysis in Section 3.3.3 works similarly as above. Of course, assumption **(A.4)*** is now too strong, and we have to dismiss it.

However, it is still possible to get a result analogous to Theorem 3.4. More precisely, we have the following statement for the unconstrained problem.

PROPOSITION 3.4. *Under assumptions **(A.1)-(A.2)-(A.3)**, suppose also that both $m \geq 2$ and $k \geq k_0+2$. In addition, take $\delta = \nu = 0$ in (3.24), and $\theta(x) = \varphi(x) = |x|^2$. Finally, define*

$$\widetilde{\mathcal{K}} \, := \, C\,(1+T) \, \Big\| \big(1 + |x|\big)^{-k_0+2} \Big\|_{L^2} \exp\Big(C \Big(\|\nabla a_0\|_{L^1_T(C^2_b)} + T \max\Big\{ \big|u^a\big| , \, \big|u^b\big| \Big\} \Big) \Big) \times$$
$$\times \Big(\|f_0\|_{H^2_{k_0+2}} \, + \, \|g\|_{L^1_T(H^2_{k_0+2})} \Big) ,$$

where the constant $C > 0$ is a suitable positive constant.

If the condition $\widetilde{\mathcal{K}}\, T/\gamma \, < \, 1$ holds true, then there exists at most one optimal control u^* in $\operatorname{int} U_{ad}$.

PROOF. The proof is very similar to the one to Theorem 3.4, therefore we limit ourselves to put in evidence the main changes to be adopted, and to treat the most delicate points of the analysis.

As before, let (u_1, f_1, q_1) and (u_2, f_2, q_2) be two optimal controls with corresponding state and adjoint state. Arguing as above, we find that $\delta u = u_1 - u_2$ fulfils estimate (3.42). Let us now focus on the estimate of each integral appearing in that relation. As for the former integral term, also by use of Lemma 3.2, we can write

$$\int_{\mathbb{R}^d} \left|\operatorname{div}\Big((e+x)\,\delta f(t)\Big)\,q_1(t)\right|\,dx \,\leq\, \|q_1(t)\|_{L^2_{-k_0}}\,\left\|\operatorname{div}\Big((e+x)\,\delta f(t)\Big)\right\|_{L^2_{k_0}} \leq\, C_3\,\|\delta f(t)\|_{H^1_{k_0+1}}\,.$$

Notice that the constant C_3 can be expressed as follows

$$(3.49)\quad C_3 \,:=\, C\,(1+T)\,\left\||x|^2\Big(1+|x|\Big)^{-k_0}\right\|_{L^2}\,\exp\Big(C\,\Big(\|\operatorname{div} a_0\|_{L^1_T(L^\infty)} + \|u_1\|_{\mathbb{L}^1_T}\Big)\Big)\,,$$

for a "universal" constant $C > 0$. At this point, the estimate for δf works as before, finally leading to

$$(3.50)\qquad \int_{\mathbb{R}^d} \left|\operatorname{div}\Big((e+x)\,\delta f(t)\Big)q_1(t)\right|\,dx \,\leq\, \widetilde{C}_3\,\int_0^t \left|\delta u(s)\right|ds\,,$$

where we have defined $\widetilde{C}_3 \,=\, C_3\,K_1^{(k_0+2)}$, just depending on the data of the problem. Next, consider the second integral in (3.42): we can estimate

$$\int_{\mathbb{R}^d} \left|\operatorname{div}\Big((e+x)\,f_2(t)\Big)\,\delta q(t)\right|\,dx \,\leq\, \|\delta q(t)\|_{L^2_{-k_0}}\,\|f_2(t)\|_{H^1_{k_0+1}} \,\leq\, C_4\,\|\delta q(t)\|_{L^2_{-k_0}}\,,$$

where, by Theorem 2.4 applied to equation (3.31) for f_2, we obtain that

$$(3.51)\quad C_4 \,:=\, C\,\exp\Big(C\,\Big(\|\nabla a_0\|_{L^1_T(C^1_b)} + \|u_2\|_{\mathbb{L}^1_T}\Big)\Big)\,\Big(\|f_0\|_{H^1_{k_0+1}} + \|g\|_{L^1_T(H^1_{k_0+1})}\Big)\,,$$

for a "universal" constant $C > 0$. On the other hand, Lemma 3.2 applied to the equation for δq gives, for a new constant $C > 0$, the estimate

$$\|\delta q(t)\|_{L^2_{-k_0}} \,\leq\, C\,\exp\Big(C\,\Big(\|\operatorname{div} a_0\|_{L^1_T(L^\infty)} + \|u_1\|_{\mathbb{L}^1_T}\Big)\Big)\times \\ \times \int_t^T |\delta u(s)|\,\left\|\Big(1+|x|\Big)\,\nabla q_2(s)\right\|_{L^2_{-k_0}}\,ds\,.$$

Notice that $\left\|\Big(1+|x|\Big)\,\nabla q_2(s)\right\|_{L^2_{-k_0}} \,\leq\, \|\nabla q_2(s)\|_{L^2_{-k_0+1}}$. In order to bound this quantity, we can differentiate the equation for q_2 with respect to x^j, for $1 \leq j \leq d$, and get (notice that $\partial_j |x|^2 \,=\, 2\,x^j$)

$$\partial_t\left(\Big(1+|x|\Big)^{-k_0+1}\,\partial_j q_2\right) \,+\, a(t,x;u_2)\cdot\nabla\left(\Big(1+|x|\Big)^{-k_0+1}\,\partial_j q_2\right) \,= \\ =\, 2\,x^j\,\Big(1+|x|\Big)^{-k_0+1} \,-\, \Big(1+|x|\Big)^{-k_0+1}\,\partial_j a(t,x;u_2)\cdot\nabla q_2\,,$$

with initial datum equal to $2\,x^j\,\Big(1+|x|\Big)^{-k_0+1}$. Obviously, the latter term in the right-hand side can be absorbed by a Grönwall argument; in addition, an easy computation shows that the former is in L^2. Therefore, by applying an L^2 estimate of Theorem 2.3 to the previous equation implies, for a "universal" constant $C > 0$, the following bound:

$$\|\nabla q_2(s)\|_{L^2_{-k_0+1}} \leq C\,\exp\Big(C\,\big(\|\nabla a_0\|_{L^1_T(L^\infty)} + \|u_2\|_{\mathbb{L}^1_T}\big)\Big)\,(1+T)\,\Big\||x|\,(1+|x|)^{-k_0+1}\Big\|_{L^2}\,.$$

By use of this latter estimate, we finally obtain

$$\int_{\mathbb{R}^d}\Big|\mathrm{div}\,\Big((e+x)\,f_2(t)\Big)\,\delta q(t)\Big|\,dx\;\leq\;\widetilde{C}_4\int_t^T\Big|\delta u(s)\Big|\,ds\,, \tag{3.52}$$

where we have defined $\widetilde{C}_4\;:=\;C_4\,\widetilde{\mathcal{K}}_1^{(1)}$ and

$$\widetilde{\mathcal{K}}_1^{(1)}\;:=\;C\,\exp\Big(C\,\big(\|\nabla a_0\|_{L^1_T(C^1_b)} + \|u_1\|_{\mathbb{L}^1_T} + \|u_2\|_{\mathbb{L}^1_T}\big)\Big)\,(1+T)\,\Big\||x|\,(1+|x|)^{-k_0+1}\Big\|_{L^2}\,. \tag{3.53}$$

We can now insert (3.50) and (3.52) into (3.42), and conclude as done in the proof to Theorem 3.4. □

3.4. Numerical analysis of ensemble control problems

The last fundamental step in solving our ensemble optimal control problems is the design of a numerical optimization procedure. For this purpose, one recognizes that the optimality condition equation (3.37) provides the semi-smooth gradient of the ensemble-cost functional along the constraint given by the Liouville model. However, because of the presence of control constraints and the combination of L^2-, L^1- and H^1-costs, the assembling of our gradient is challenging. In particular, by imposing constraints on the value of the control, we are required to implement a H^1 projection of the control update. At this point, we remark that the combination of L^1- and H^1-costs and the H^1 projection are less investigated in the literature.

However, this effort is very well justified by our purpose of implementing a state-of-the-art semi-smooth Krylov-Newton methodology for our new class of PDE optimal control problems. In doing this, we also rely on the results in [**48, 49**], and the resulting Newton scheme is used to validate our optimal control framework.

Further, we notice that the optimality condition equation is a variational inequality involving an integral for which we use second-order accurate quadratures, and we implement a projection step in the optimization procedure.

Notice that, while second-order accuracy for the above three components of the optimality system (3.31)–(3.33) is separately guaranteed by suitable approximation, we are not able to prove this order of accuracy of the coupled system; this is an issue that remains widely open in the scientific literature, apart of the case of much simpler problems with linear control mechanisms; see, e.g., [**27**].

Section 3.4.1 is devoted to the implementation of our semi-smooth Krylov-Newton method that requires the numerical solution of the Liouville equation and its adjoint and the implementation of the gradient together with a H^1-projection procedure for the controls.

In Section 3.4.2, we present results of numerical experiments with our solution methodology that validate our optimal control framework in terms of the ability of the controls to perform the given tasks. For this purpose, we consider the tracking of non-differentiable trajectories and also the case of bimodal distributions.

3.4.1. A projected semi-smooth Krylov-Newton method

In this section, we illustrate a semi-smooth Krylov-Newton (SSKN) method for solving the ensemble optimal control problem (3.5)–(3.6) with the drift given by (2.5) and the cost functional setting specified in Section 3.1. We remark that our SSKN scheme belongs to the class of projected semi-smooth Newton schemes discussed in [**127**].

In general, a Newton method is an iterative procedure aiming at finding roots of a given function. Its peculiarity is that it may generate a sequence that can converge superlinearly or even quadratically to the sought solution.

In order to explain the Newton method in simple terms, consider the problem to find a root $\zeta^\star \in \mathbb{R}^N$ of a map $\mathcal{M} : \mathbb{R}^N \to \mathbb{R}^N$, as follows

$$\mathcal{M}(\zeta^\star) = 0, \tag{3.54}$$

where, for the moment, we assume that $\zeta \mapsto \mathcal{M}(\zeta)$ is continuously differentiable.

Now, denote with $\mathcal{J}(\zeta)$ the Jacobian of $\mathcal{M}$ at ζ. The Newton method generates a sequence $(\zeta^\ell)_\ell$ by means of the following two steps

$$\begin{aligned} \mathrm{S}_1: &\quad \Delta\zeta^\ell = -(\mathcal{J}(\zeta^\ell))^{-1}\,\mathcal{M}(\zeta^\ell) \\ \mathrm{S}_2: &\quad \zeta^{\ell+1} = \zeta^\ell + \Delta\,\zeta^\ell. \end{aligned} \tag{3.55}$$

The steps S_1–S_2 are performed for $\ell = 0, 1, 2, \ldots$, starting with a given initial guess ζ^0.

Clearly, the Newton sequence is well defined if the Jacobian is invertible at each iterate, and we assume that this is the case in a neighbourhood $\mathcal{N}$ of the solution $\zeta^\star$, where also the inverse is uniformly bounded. With these assumptions and requiring that the initial guess $\zeta^0 \in \mathcal{N}$ is sufficiently close to $\zeta^\star$, one can prove that the sequence (ζ^ℓ) converges quadratically to the root $\zeta^\star$, that is, $\|\zeta^{\ell+1}-\zeta^\star\|_2 \le c\,\|\zeta^\ell-\zeta^\star\|_2^2$, for some constant $c > 0$, and $\|\cdot\|_2$ denotes the Euclidean norm of a vector in $\mathbb{R}^N$. However, in the case where $\mathcal{M}$ is only differentiable and provided that the following holds

$$\|\mathcal{M}(\zeta + \delta\zeta) - \mathcal{M}(\zeta) - \mathcal{J}(\zeta)(\delta\zeta)\|_2 = o(\|\delta\zeta\|_2) \quad \text{as} \quad \delta\zeta \to 0, \tag{3.56}$$

then the Newton sequence converges at least superlinearly, i.e., faster than linearly.

The same Newton procedure (3.55) can be applied to find an extremal of the minimization problem $\min_{\zeta\in\mathbb{R}^N} f(\zeta)$, by considering $\mathcal{M}(\zeta) := \nabla f(\zeta)$, where ∇ denotes the gradient in $\mathbb{R}^N$, and assuming that $f : \mathbb{R}^N \to \mathbb{R}$ is twice differentiable. In this case, we have that $\mathcal{J}(\zeta) = \nabla^2 f(\zeta)$.

Now, in the case of a constrained optimization problem $\min_{\zeta\in K} f(\zeta)$, where $K \subset \mathbb{R}^N$ is closed and convex, an extremal $\zeta^\star$ of this problem is characterized by the inequality $\nabla f(\zeta^\star)\cdot(\zeta-\zeta^\star) \geq 0$. However, this inequality can be equivalently written as follows

$$\mathcal{F}(\zeta^\star) := \zeta^\star - P_K\left(\zeta^\star - \mathfrak{s}\,\nabla f(\zeta^\star)\right) = 0, \tag{3.57}$$

where P_K is the projection of $\mathbb{R}^N$ onto K, and $\mathfrak{s} > 0$ is arbitrary but fixed. Therefore, the solution of the optimality condition in the form of an inequality can be reformulated as a root problem. However, even if f is continuously differentiable, the function $\mathcal{F}$ is not. On the other hand, if ∇f is locally Lipschitz, then also $\mathcal{F}$ is locally Lipschitz continuous.

We see that a lack of differentiability of $\nabla f(\zeta)$ or the presence of constraints as above hinder the application of the Newton scheme to solve optimization problems. This situation has motivated a great effort towards the generalization of the notion of differentiability that makes possible to pursue the Newton approach also in non-differentiable cases; see [**52, 51, 127**] for details and further references.

The main assumption for this generalization is the Lipschitz continuity of the map $\zeta \mapsto \mathcal{F}(\zeta)$, in which case Rademacher's theorem [**127**] states that this map is almost everywhere differentiable. Based on this result, the notion of differentiability has been extended as follows; see [**127**] for a detailed discussion.

Definition 3.1. *Assuming $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^N$ be locally Lipschitz continuous, we have the following generalized Jacobians of $\mathcal{F}$ at ζ:*

(a) *the Bouligand subdifferential given by*

$$\partial_B\mathcal{F}(\zeta) := \Big\{S \in \mathbb{R}^{N\times N} \,:\, \exists\{\zeta^\ell\}_\ell \subset \mathbb{R}^N \setminus U_{nd} \,:\, \zeta^\ell \to \zeta\,,\ \mathcal{J}(\zeta^\ell) \to S\Big\},$$

where U_{nd} is the set of points where $\mathcal{F}$ fails to be Fréchet differentiable and $\mathcal{J}(\zeta)$ denotes the Jacobian of $\mathcal{F}$ at ζ;

(b) *the Clarke's subdifferential is the convex hull of $\partial_B\mathcal{F}(\zeta)$, denoted with $\partial\mathcal{F}(\zeta) := \mathrm{co}\ \partial_B\mathcal{F}(\zeta)$.*

With this construction, we can apply (3.55) by choosing a generalized Jacobian $\widetilde{\mathcal{J}}^\ell \in \partial\mathcal{F}(\zeta^\ell)$. However, in order to guarantee superlinear convergence of the resulting Newton sequence, the following property of semi-smoothness is required; see [**108, 127**].

Definition 3.2. *A locally Lipschitz continuous function $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^N$ is said to be semi-smooth at $\zeta \in \mathbb{R}^N$ if and only if $\mathcal{F}$ is directionally differentiable at ζ, and it satisfies the condition*

$$\max_{\widetilde{\mathcal{J}} \in \partial \mathcal{F}(\zeta+\delta\zeta)} \|\mathcal{F}(\zeta+\delta\zeta) - \mathcal{F}(\zeta) - \widetilde{\mathcal{J}}\,(\delta\zeta)\|_2 = o(\|\delta\zeta\|_2) \quad \textit{as} \quad \delta\zeta \to 0. \tag{3.58}$$

Notice that the discussion above has focused on finite dimensional spaces, which is also the case of our numerical optimization problem. However, the subdifferential framework given above has been extended also to maps acting between infinite-dimensional Banach spaces [**52, 51, 127**]. In particular, we can apply it to our ensemble optimal control problem (3.29), that is, $\min_{u\in U_{ad}} J_r(u) := J\Big(G(u), u\Big)$. One can recognize that $J_r(u)$ is not Fréchet differentiable due to the presence of the L^1-cost, which however is Lipschitz in u. In fact, in Section 3.3, we have used sub-differential calculus [**127**] to determine the gradient $\widetilde{\nabla} J_r(u)$, and formulated the first-order optimality condition $(\widetilde{\nabla} J_r(u), v-u)_U \geq 0$ for all $v \in U_{ad}$. Therefore, we can proceed as in (3.57) and consider the application of the Newton scheme with generalized Jacobian to the equation

$$u - P_{U_{ad}}\,(u - \mathfrak{s}\,\widetilde{\nabla} J_r(u)) = 0.$$

However, although this procedure is standard with control problems with L^2-L^1 costs [**49, 48, 127**], it becomes very cumbersome in our case with H^1 costs. For this reason, we consider a projected semi-smooth Newton (pSSN) scheme with the following steps

$$\begin{aligned} \mathrm{S}_1 : \quad & \Delta u^\ell = -(\widetilde{\mathcal{J}}(u^\ell))^{-1}\,\widetilde{\nabla} J_r(u^\ell) \\ \mathrm{S}_2 : \quad & u^{\ell+1} = P_{U_{ad}}\,\left(u^\ell + \mathfrak{s}\,\Delta\, u^\ell\right). \end{aligned} \tag{3.59}$$

with $\ell = 0, 1, 2, \ldots$, and starting with a given initial guess $u^0 \in U_{ad}$. In the following, we discuss the step S_1 and thereafter S_2.

Concerning the step S_1, we see that the main computational effort in the procedure (3.59) would be the assembly and inversion of the Jacobian $\widetilde{\mathcal{J}}(u^\ell)$, but this is not possible because of the size of the problem. In fact, in PDE optimization, one implements the action of the Jacobian (reduced Hessian) on a vector and uses a Krylov approach. Thus, we replace the step S_1 in this procedure with the step: Solve

$$\widetilde{\mathcal{J}}(u^\ell)\,\Delta u^\ell = -\,\widetilde{\nabla} J_r(u^\ell)$$

by a Krylov method (e.g., minres) to a given tolerance. In this way, we have a projected SSKN scheme.

Next, we illustrate how the action of the Jacobian on the increment Δu^ℓ is constructed. For this purpose, we determine the second-order directional derivative of our Lagrange functional (3.30) with respect to u, and this requires to consider the linearizations of the forward and adjoint equations with respect to u. We have the following

$$\widetilde{\mathcal{J}}(u)\,\Delta u = \Big(\nabla_{uu}\mathcal{L}\Big)(\Delta u) + \Big(\nabla_{uf}\mathcal{L}\Big)(\hat{f}) + \Big(\nabla_u L\Big)^*(\hat{q}), \tag{3.60}$$

where $L(f,u) := \partial_t f + \operatorname{div}(a(u)\,f)$ represents the Liouville operator, and $*$ means adjoint. In (3.60), the function $\hat{f}$ is the solution of the following linearized Liouville problem

$$\partial_t \hat{f} + \operatorname{div}(a\,\hat{f}) = -\operatorname{div}(\hat{a}\,f), \qquad \text{with} \qquad \hat{f}_{|t=0} = 0, \tag{3.61}$$

where $\hat{a} = \frac{\partial a}{\partial u}\,\Delta u$.

Equation (3.61) is obtained in the following way. First, define a small variation $\hat{f}$ of f such that $(f+\hat{f}) \in C_T(L^2(\mathbb{R}^d))$ and $(\hat{f})_{|t=0} = 0$. Then, insert $f+\hat{f}$ for f in equation (3.6), use the linearity of (2.1) with respect to f and take into account that f itself solves (2.1).

To complete the discussion of (3.60), we explain how to compute $\hat{q}$. It is obtained solving the following linearized adjoint problem, resulting from a linearization procedure similar to that for $\hat{f}$. We have

$$-\partial_t \hat{q} - a \cdot \nabla \hat{q} = \hat{a} \cdot \nabla q, \qquad \text{with} \qquad \hat{q}_{|t=T} = 0. \tag{3.62}$$

We solve (3.61) and (3.62) with our KTS scheme. For further details on the implementation of the action of the Jacobian on a vector, we refer to, e.g., [**26**], Chapter 6.3.5.

Specifically, for our case one can verify that (3.60) is explicitly given component-wise by

$$\Big(\widetilde{\mathcal{J}}(u)(\Delta u)\Big)_{m,r} := (\Delta u)^r_m + \Phi_{m,r}, \qquad m = 1,2, \qquad r = 1,2,$$

where the components of Φ are solutions to the following boundary-value problem

$$\Big(-\nu\frac{d^2}{dt^2} + \gamma\Big)\,\Phi_{m,r} = -\int_{\mathbb{R}^2} \frac{\partial a}{\partial u^r_m}\hat{f} \cdot \nabla q \; dx + \int_{\mathbb{R}^2} \operatorname{div}\Big(\frac{\partial a}{\partial u^r_m} f\Big)\,\hat{q}\; dx$$
$$\Phi_{m,r}(0) = 0, \qquad \Phi_{m,r}(T) = 0.$$

We approximate this problem by finite differences for the time derivative, which results in a tridiagonal linear system, and solve it with the Thomas algorithm. Compare this boundary-value problem with (3.41).

Notice that, in general in optimal control problems, the reduced Hessian has a favourable spectral structure, in the sense that it is spectrally equivalent to a second-kind Fredholm integral operator, and in this case Krylov solvers can converge in a mesh-independent number of iterations [**3**]. This is supported by our estimate of the condition number of the Jacobian. For this purpose, we use the power method to estimate the largest eigenvalue of $\widetilde{\mathcal{J}}(u)$, and the inverse power method to estimate its smallest eigenvalue, and obtain an approximation to the condition number of the Jacobian; here we assume, that we can approximate the condition number by the ratio of the approximation of the maximal and the minimal eigenvalue of $\widetilde{\mathcal{J}}(u)$. Notice that these methods also do not require the assembly of $\widetilde{\mathcal{J}}(u)$. We compute this condition number in correspondence to different mesh sizes and report these values in Table 3.1. A plot of these values for different mesh sizes is shown in Figure 3.1, where we see that the condition number is of the same order of magnitude for different mesh sizes.

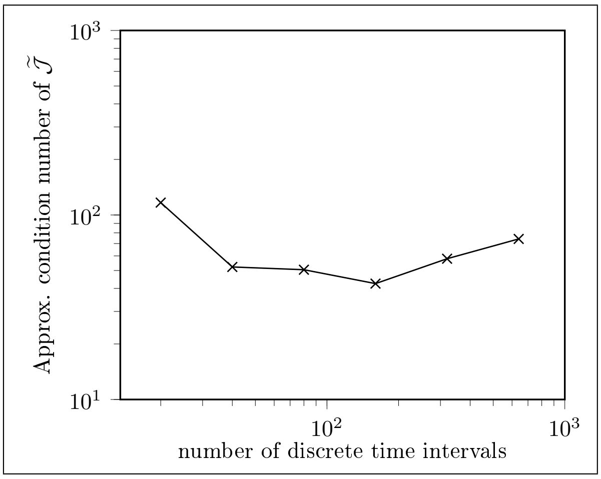


Figure 3.1. Approx. condition number of $\widetilde{\mathcal{J}}$

N_t	$c(\widetilde{\mathcal{J}})$
20	116.70
40	52.27
80	50.51
160	42.47
320	57.94
640	74.18

Table 3.1. Approx. condition number $c(\widetilde{\mathcal{J}})$

Next, we discuss step S_2 of (3.59). This step is required to ensure that any control update results in a control function in U_{ad}. To implement the H^1 projection, denoted with $P_{U_{ad}}$, we solve the following optimization problem

$$\min_{\tilde{u}\in U_{ad}} \frac{1}{2} \|\tilde{u} - u\|^2_{\widetilde{\mathbb{H}}^1_T} . \tag{3.63}$$

Since $\widetilde{\mathbb{H}}^1_T$ is a Hilbert space and U_{ad} is non-empty, closed and convex, we know that there exists a unique projection (see [**112**], Theorem 4.11). The problem (3.63) can equivalently be written as follows

$$\begin{cases} \min_{\tilde{u}\in\mathbb{H}^1_T} J_P(u) := \frac{1}{2}\|\tilde{u}-u\|^2_{L^2_T} + \frac{1}{2}\left\|\frac{d}{dt}(\tilde{u}-u)\right\|^2_{L^2_T} \\ \text{s.t.} \qquad \max(u^a - \tilde{u}) = 0, \qquad \max(\tilde{u} - u^b) = 0. \end{cases} \tag{3.64}$$

Notice that, corresponding to this optimization problem, we have the following Lagrange functional with Lagrange multipliers q_a, q_b,

$$\mathrm{l}(u, q_a, q_b) := J_P(u) + \int_0^T \max(u - u^b, 0)\, q_b\, dt + \int_0^T \max(u^a - u, 0)\, q_a\, dt$$

To solve this optimization problem to implement the projection $P_{U_{ad}}$, we use a gradient descent scheme; see, e.g., [**19**], Section 2.8.

Next, we summarize the Newton procedure to solve our Liouville ensemble optimal control problem in the following algorithm that determines the reduced gradient at a given u.

Algorithm 3.1 Computation of the gradient $\widetilde{\nabla} J_r(u)$

```
Require: $u$
Solve the Liouville equation (3.31) with Algorithm 2.1
Solve the adjoint Liouville equation (3.32) with Algorithm 2.2
Assemble the $L^2$ gradient in (3.39)
Assemble the $H^1$ gradient given by $\widetilde{\nabla} J_r(u)$ in (3.36)
return $\widetilde{\nabla} J_r(u)$
```

With this algorithm, we can define our projected semi-smooth Krylov-Newton algorithm as follows.

Algorithm 3.2 Projected semi-smooth Krylov-Newton method

```
Require: $u^0$
Set $\ell = 0$, $E > tol$
while $E > tol$ and $\ell < \ell_{\max}$ do
    Compute $\widetilde{\nabla} J_r(u^\ell)$ with Algorithm 3.1
    Solve $\widetilde{\mathcal{J}}(u^\ell)\,\Delta u = -\widetilde{\nabla} J_r(u^\ell)$ (we use minres; here we need to solve (3.61) and (3.62))
    Set $u^{\ell+1} = P_{U_{ad}}\left(u^\ell + \mathfrak{s}\,\Delta u\right)$, where $\mathfrak{s}$ is determined by the Armijo linesearch-backtracking scheme.
    Set $E = \left\| u^{\ell+1} - u^\ell \right\|_{1,h}$
    $\ell = \ell + 1$
end while
Solve the Liouville equation (3.31) with Algorithm 2.1
return $(f(u^\ell), u^\ell)$
```

In this algorithm, we use the difference between consecutive iterations of the control as termination criterion, specifically to stop the algorithm, if the difference is less then a threshold $tol > 0$. Moreover, we define a maximum number of iterations $\ell_{\max} \in \mathbb{N}$.

3.4.2. Numerical experiments

Next, we validate the ability of our optimization framework to construct controls that steer the ensemble density in order to follow a desired path. For this purpose, we start considering the tracking of a piecewise smooth trajectory with an initial density given by a unimodal distribution. Thereafter, we demonstrate that our approach allows to construct control functions that are able to drive the evolution of the density with a bimodal structure.

In our first experiment on tracking, we choose $\Omega = [-1,1] \times [-1,1]$, and the initial density on this domain is given by

$$f_0(x) := \frac{C_0}{2\pi\sigma^2} \exp\left(-\frac{|x-\xi_0|^2}{2\sigma^2}\right).$$

This is a unimodal Gaussian distribution centred in $\xi_0 = (-0.5, 0.5)$, with variance $\sigma = \frac{1}{4}$, and we take $C_0 = \frac{1}{10}$. Notice that, by this choice, the value of f_0 at the boundary of Ω is of the order of machine precision, and further it holds that $f_0 \in C^\infty(\Omega)$.

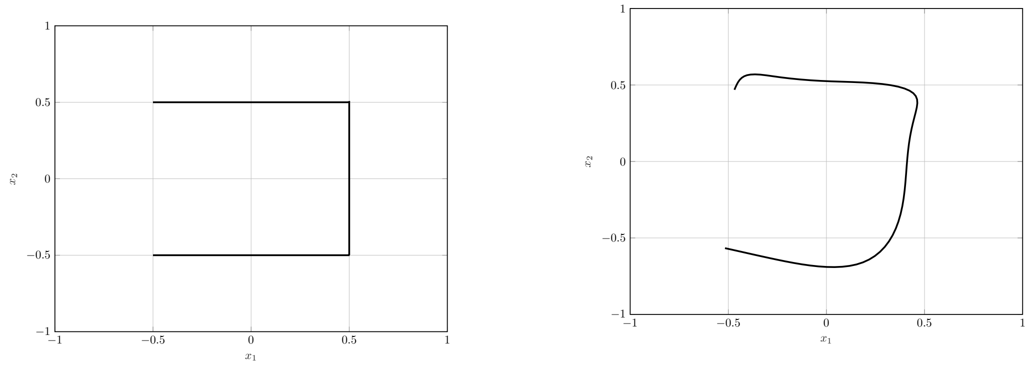
In this experiment, the purpose of the control is to drive the ensemble of trajectories along the following piecewise smooth desired trajectory

$$\xi_D(t) := \begin{cases} \left(\frac{3t}{T} - \frac{1}{2}, \frac{1}{2}\right) & 0 \le t \le \frac{T}{3} \\ \left(\frac{1}{2}, \frac{3}{2} - \frac{3t}{T}\right) & \frac{T}{3} < t \le \frac{2T}{3} \\ \left(\frac{5}{2} - \frac{3t}{T}, -\frac{1}{2}\right) & \frac{2T}{3} < t \le T. \end{cases} \tag{3.65}$$

A plot of this trajectory in Ω is given in Figure 3.2a. Correspondingly, our potentials in the objective functional are chosen as follows

$$\theta(x,t) = -\frac{C_\theta}{2\pi\sigma_\theta^2} \exp\left(-\frac{|x-\xi_D(t)|^2}{2\sigma_\theta^2}\right), \qquad \varphi(x) = -\frac{C_\varphi}{2\pi\sigma_\varphi^2} \exp\left(-\frac{|x-\xi_D(T)|^2}{2\sigma_\varphi^2}\right),$$

where $x \in \Omega$ and $t \in [0,T]$, and the values of C_θ, C_φ, σ_θ and σ_φ are given in Table 3.2.



(a) Parametric plot of ξ_D over time.

(b) Parametric plot of the mean $\mathbb{E}[x]$ over time.

Figure 3.2. Setting and results of the first experiment.

Param.	Value	Param.	Value
γ	$5 \cdot 10^{-4}$	δ	10^{-4}
ν	$5 \cdot 10^{-4}$	N_x	26
N_t	80	T	3
$u_{\max}$	1.5	$u_{\min}$	-1.5
C_θ	10	σ_θ	0.45
C_φ	$C_\theta \frac{T}{N_t-1}$	σ_φ	0.45

Table 3.2. Parameters' setting for the first experiment.

Param.	Value	Param.	Value
γ	10^{-4}	δ	10^{-5}
ν	10^{-4}	N_x	51
N_t	150	T	3
$u_{\max}$	1	$u_{\min}$	-1
C_θ	10	σ_φ	0.45
C_φ	$C_\theta \frac{T}{N_t-1}$	σ_φ	0.45

Table 3.3. Parameters' setting for the second experiment.

Now, we specify a setting that facilitates a comparison of our results of ensemble control with a simple dynamic for the trajectory. Specifically, suppose that our desired trajectory is the result of the following dynamics

$$\dot{\xi}_D(t) = u_1(t), \qquad \xi_D(0) = \xi_0. \tag{3.66}$$

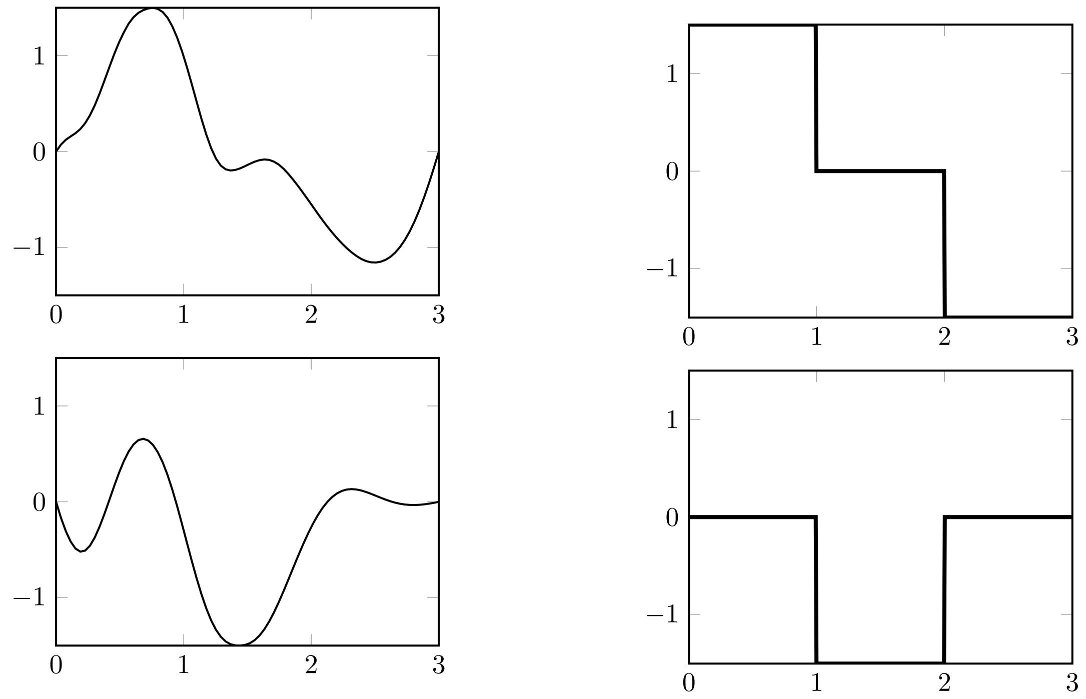
Then we can immediately compute the control u_1 in this equation such that the solution to (3.66) is given by (3.65). This control is plotted in Figure 3.3b, and we refer to it as the single-trajectory control, specifically taken $u_2 \equiv 0$, which corresponds to no change in the variance. Notice that this control is not in our control space U_{ad} (recall its definition (3.23) above) since it is not continuous. Moreover, in its construction, we do not require to satisfy the conditions of its value being zero at initial and final times.

In our drift (2.5), we choose $a_0 = 0$, $a_1 = 1$ and $a_2 = 0$, and with this setting we solve our Liouville control problem, taking the numerical values given in Table 3.2. The resulting control function is depicted in Figure 3.3a, which appears similar to the single-trajectory control in Figure 3.3b. We see that the former is in $\mathbb{H}^1_T$ and is zero at $t = 0$ and $t = T$ as required, we refer to it as the Liouville control.

Corresponding to the Liouville control, we obtain an evolution of the density with which we compute the function $\mathbb{E}[x](t) = \int x\, f(x,t)\, dx$. This function is shown in Figure 3.2b. Notice that it closely resembles the desired trajectory.

In our second experiment, we consider the setting $a_0 = 0$, $a_1 = 1$ and $a_2 = 1$, and we take a smooth initial f_0 that is given by a bimodal Gaussian distribution as follows

$$f_0(x) = \frac{C_0}{2\pi\sigma^2} \exp\left(-\frac{|x-\xi_0^1|^2}{2\sigma^2}\right) + \frac{C_0}{2\pi\sigma^2} \exp\left(-\frac{|x-\xi_0^2|^2}{2\sigma^2}\right), \tag{3.67}$$



(a) Liouville control over time; top: u_1^1, bottom: u_1^2

(b) Single-trajectory control over time; top: u_1^1, bottom: u_1^2

Figure 3.3. Comparison of controls for the first experiment.

where

$$\xi_0^1 = \left(-\frac{3}{4}, \frac{3}{4}\right), \qquad \xi_0^2 = \left(-\frac{3}{4}, -\frac{3}{4}\right), \qquad \sigma = \frac{1}{4}, \quad C_0 = \frac{1}{10}.$$

The values of the other parameters are specified in Table 3.3.

In this case, we choose the following desired trajectory

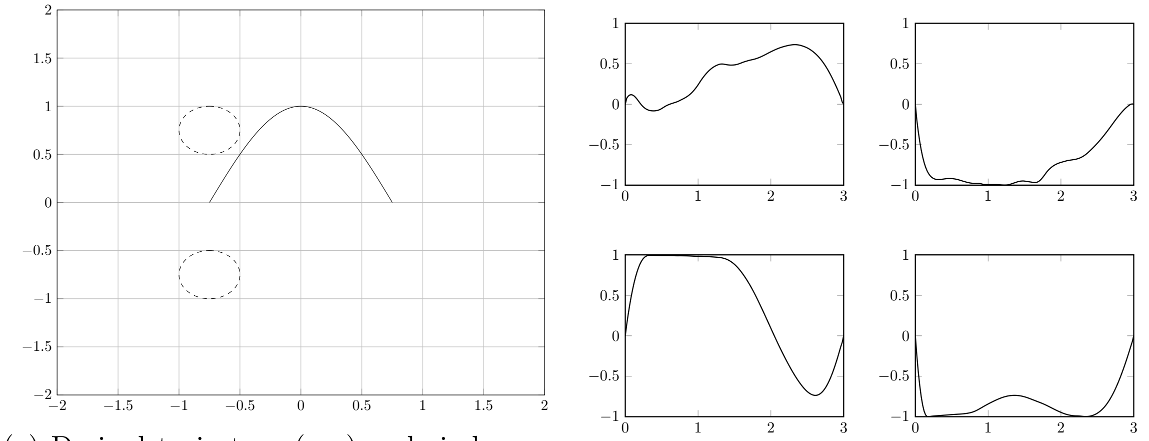
$$\xi_D(t) = \left(-\frac{3}{4} + \frac{3t}{2T}, \sin\left(\frac{\pi t}{T}\right)\right).$$

We have that $\xi_D(0)$ corresponds to the midpoint between the centres of the two Gaussians defining the initial density; see Figure 3.4a, where we plot circles around the centres of the two Gaussians with radius of their standard deviation.

With this setting, we solve our Liouville optimal control problem and obtain the controls shown in Figure 3.4b. The values of C_θ, C_φ, σ_θ and σ_φ together with the values of the numerical parameters are given in Table 3.3.

Corresponding to these controls, we obtain the evolution of the density depicted in Figure 3.5a. Specifically, we plot the shape of the density f at all times. One can see that the bimodal density is driven towards the desired trajectory becoming unimodal. The same result is visualized in Figure 3.5b from a different perspective.

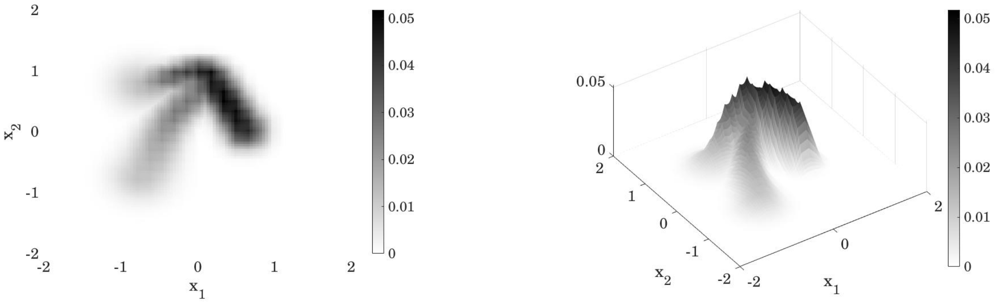
We would like to conclude this section considering a setting that generalizes our framework. Our purpose is to demonstrate that our control framework is also able



(a) Desired trajectory (——) and circles centred at ξ_0^1 and ξ_0^2 with radius σ identifying the Gaussians (- - -)

(b) Optimal controls; top left: u_1^1, top right: u_1^2, bottom right: u_2^1, bottom left: u_2^2

Figure 3.4. Setting and results of the second experiment.



(a) Evolution of f over time; top view.

(b) Evolution of f over time; side view.

Figure 3.5. Evolution of the density in the second experiment.

to drive a smooth bimodal distribution to follow two trajectories. In this case, we choose a potential θ that resembles a double well, so that it provides two basins of attraction corresponding to the two trajectories.

In this experiment, we consider the initial bimodal f_0 given in (3.67), and consider the following two desired trajectories

$$\xi_{D1}(t) = \left(-\frac{3}{4} + \frac{3t}{2T}, \frac{3}{4} - \frac{3t}{4T}\right)^T, \qquad \xi_{D2}(t) = \left(-\frac{3}{4} + \frac{3t}{2T}, -\frac{3}{4} + \frac{3t}{4T}\right)^T$$

In correspondence to these trajectories, we define θ as follows

$$\theta(x,t) = -\frac{C_\theta}{2\pi\sigma_\theta^2}\left[\exp\left(-\frac{|x-\xi_{D1}(t)|^2}{2\sigma_\theta^2}\right) + \exp\left(-\frac{|x-\xi_{D2}(t)|^2}{2\sigma_\theta^2}\right)\right].$$

Similarly, we define $\varphi(x) = \theta(x,T)$.

We solve the resulting ensemble control problem with Algorithm 3.2 and obtain the controls depicted in Figure 3.6. In correspondence to these controls, we obtain the

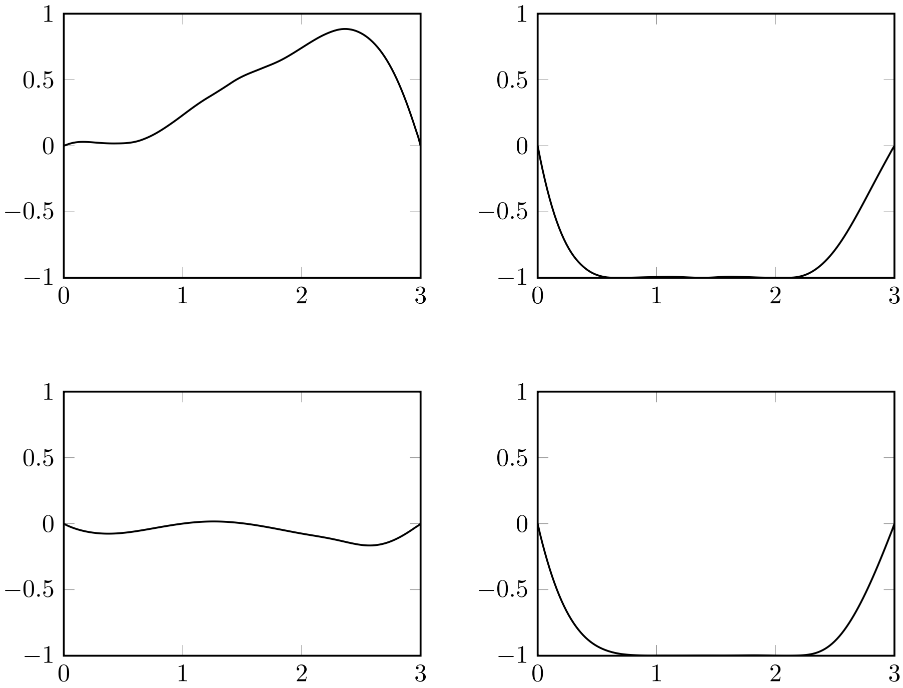
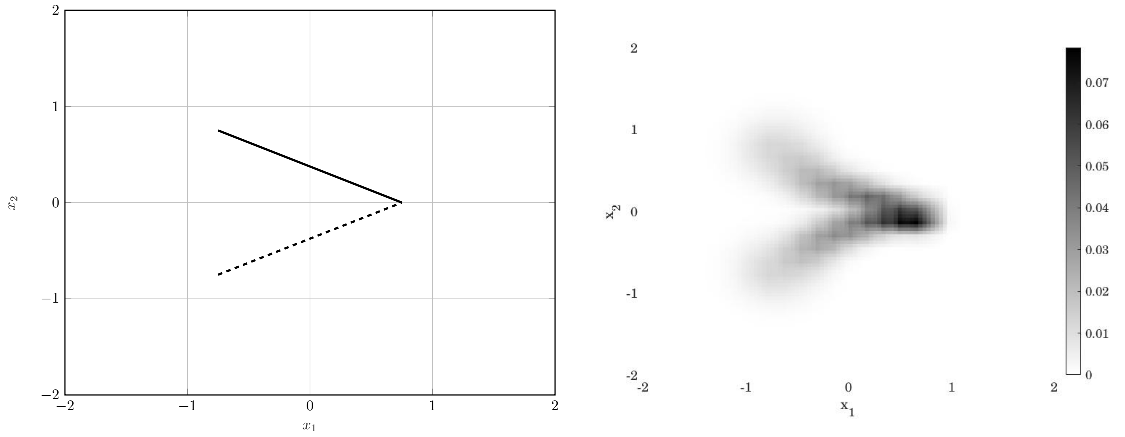


Figure 3.6. Optimal controls in the third experiment; top left: u_1^1, top right: u_2^1, bottom right: u_2^1, bottom left: u_2^2.



(a) Desired trajectories ξ_{D1} (solid line) and ξ_{D2} (dashed line).

(b) Evolution of the density over time.

Figure 3.7. Results of the third experiment.

evolution of the initial bimodal density shown in Figure 3.7b. We see that the two initial Gaussians are driven along the desired trajectories ξ_{D1} and ξ_{D2} shown in Figure 3.7a and merge at the final time.

These experiments finish the investigation of kinetic models without collisions. In the next chapter, we turn to linear kinetic models and include the modelling of collisions.

Chapter 4

Linear kinetic models

In this chapter, we investigate linear kinetic models including collisions. The paramount concept of kinetic theory is that of a distribution function f of a sufficiently rarefied gas consisting of hard spherical particles defined in the phase space spanned by the position $x \in \mathbb{R}^3$ and velocity $v \in \mathbb{R}^3$. In this statistical framework, the fundamental model that governs the time evolution of f is the Boltzmann equation [**25**]. Since its appearance, this framework has been widely investigated because it also provides the intermediate (mesoscopic) step in the transition between atomistic and continuous models for gas dynamics; see the recent works [**9, 24, 107**] and the references therein.

On the other hand, the Boltzmann equation is a fundamental simulation tool in applications where the mesoscopic scale is predominant. In fact, it is employed in a wide class of problems, ranging from aerodynamics and space propulsion [**83**] to microscale electronic devices and materials [**56, 111**], ionized dilute gases [**134**] and high-temperature plasma [**84**], and other applications in multi-particle systems [**5**]. Indeed, this is a very short and incomplete list of references concerning present and envisioned application problems that often go much beyond the setting of the original formulation of the Boltzmann equation.

In this chapter, we focus on a representative model for linear kinetic theory proposed by J. Keilson and J.E. Storer [**86**], and we consider a control mechanism in the collision kernel that models the average speed of the postulated background particles. This mechanism could theoretically be realised by a temperature gradient in the fluid background [**65, 105**].

In Section 4.1, we consider the space-homogeneous case as originally proposed by Keilson and Storer and assume that the time-dependent control is within the collision kernel. We formulate an optimal control problem in this setting and derive a corresponding optimality system. Moreover, we manage to reformulate the adjoint equation in such a way, that enables us to use a fully kinetic optimization framework to solve the optimal control problem.

In Section 4.2, we extend our model to the space-inhomogeneous case and include also external forces. We formulate an optimal control problem governed by the inhomogeneous Keilson-Storer equation and consider the control mechanism as a space-dependent external force. Also for this case, we formulate a numerical optimization strategy that considers the kinetic interpretation of the arising equations. We validate our implementation by performing different numerical experiments.

4.1. The controlled Keilson-Storer collision term

We choose the so-called Keilson-Storer (KS) master equation for a multitude of reasons. Since its appearance in 1952, the KS model has been successfully utilized in a range of applications including the estimation of transport coefficients [**17**], laser spectroscopy [**16**], and molecular dynamics simulations [**125**], reorientation of molecules in liquid water [**73**], and quantum transport [**88**]. Further, the KS model is specified through a single parameter that allows to mimic strong and weak collision limits [**121**]. Moreover, recently a microscopic derivation of the KS master equation has been achieved [**72**].

The Keilson-Storer master equation was proposed in [**86**] as a variant of a linear Boltzmann equation [**42**] to model Brownian motion [**97, 99**]. This is the motion of a particle of mass M immersed in a viscous fluid that determines dynamical fluctuations, which are due to interaction of the much smaller particles of mass $m \ll M$ of the fluid with the heavier particle (the particle, in the following). If $A(w,v)\,dv$ is the probability per unit time that this particle with velocity w will undergo a transition to a ball of volume dv centred in v, then the master equation describing its motion can be written as follows [**86, 117**]

$$\partial_t f(v,t) = \Gamma \int_{\mathbb{R}^d} A(w,v)\, f(w,t)\, dw - \Gamma\, f(v,t) \int_{\mathbb{R}^d} A(v,w)\, dw, \tag{4.1}$$

where $f(v,t)$ represents the probability density of the particle to have velocity v at time t, and Γ represents the relaxation rate $1/\tau$, the inverse of the mean free time between collisions of the particle of mass M with the fluid particles. In [**86**], this rate is assumed constant. Further, in (4.1), the integrals are taken over the entire Euclidean velocity space, which we assume to be two dimensional.

Now, in order to locate the KS optimal control problems considered in this chapter in the larger context of (1.3), we remark that our KS model (4.1) is a special case of (1.3). Specifically, assuming that for the drift function it holds

$a(x, v, t; u) = (a^1(v, t; u), a^2(x, t; u))$, which is the case in the most physical applications, we can write (1.3) as

$$\partial_t f(x, v, t) + a^1 \cdot \nabla_x f(x, v, t) + a^2 \cdot \nabla_v f(x, v, t) = C[f](x, v, t).$$

Notice that the space coordinate x does not appear in the KS model because it is assumed that f is uniformly (and infinitely) distributed along the x–coordinate. This is also why the term $a^1 \cdot \nabla_x f$ does not appear in our model; see, e.g., [**43**, Chapter 6] for an analysis of this case. On the other hand, the term $a^2 \cdot \nabla_v f$ denotes the action of an external force on the particles, which can obviously play the role of a control force that we investigate in Section 4.2. In the current section, we assume $a^2 \equiv 0$. The (nonlinear) collision term $C[f]$ is the main focus of most theoretical and application works. However, already in the linear KS case, where $C[f]$ is given by the right-hand side of (4.1), we are confronted with the problem that the adjoint of $C[f]$ no longer has the structure of a master equation term. This is conceptually and practically unsatisfactory, since many simulation schemes exploit the possibility to split the transport and collision parts of the Boltzmann equation, implementing the latter based on its physical interpretation. Moreover, this physical significance may be instrumental to have more insight into a control mechanism that is included in the collision term, and in the case where a parameter identification problem for the collision is considered. In Section 4.1, we focus on a class of control problems where the control mechanism is included in the KS kernel through a control function u that is specified below. More generally, our functional setting and our optimal control formulation are similar to that discussed in the previous Chapters in the case of ensemble control problems governed by the Liouville equation. Similarly, we discuss existence of an optimal control and, subject to appropriate differentiability conditions, we derive the KS optimality system that is central in our investigation.

One can also use (4.1) to model the motion of multiple massive particles immersed in the fluid that do not interact with each other. In this case, f represents the material density that can be conveniently normalized to 1. This assumption requires that the number density of the particles is much smaller than the number density of the particles that constitute the background. In both pictures, it is required that the fluid particles are in thermal equilibrium; see, e.g., [**99**].

In this setting, Keilson and Storer suggest a structure of the kernel A that reasonably models the microscopic scattering process as a damping scheme as follows

$$A(v, w) = \mathbf{a}(w - \gamma\, v),$$

where $\mathbf{a}(\cdot)$ is a function and $0 < \gamma \lessapprox 1$ is a damping parameter. The other requirement considered in [**86**] is the so-called detailed balance that requires $A(w, v)\, f^{eq}(w) = A(v, w)\, f^{eq}(v)$, where f^{eq} denotes the equilibrium distribution that the particles will asymptotically assume. These two requirements result in the following explicit kernel

$$A(v,w) = A_0 \exp\left(-\beta\,|w-\gamma v|^2\right), \tag{4.2}$$

where $|\cdot|$ denotes the Euclidean norm in the velocity space, and A_0, β are positive constants, which are related by requiring that f^{eq} has the Maxwellian distribution and $\frac{1}{\Gamma}\int_{\mathbb{R}^d} A(v,w)\,dw = 1$.

Thus, we set $\Gamma = A_0\,(\pi/\beta)^{d/2}$, and $\beta\,(1-\gamma^2) = \frac{M}{2k_B T_p}$, where T_p denotes the temperature. Therefore, $\gamma \in (-1,1)$ and A_0 are the parameters that define the KS kernel. Furthermore, we have $f^{eq}(v) = \bar{f}_0\,\exp\left(-\beta\,(1-\gamma^2)\,|v|^2\right)$, where $\bar{f}_0$ is a normalization constant. Notice that, with $\gamma = 0$, we have $A(v,w) = f^{eq}(w)$, and the KS master equation takes the well-known BGK structure as follows [**21**]

$$\partial_t f(v,t) = \Gamma\,\left(f^{eq}(v) - f(v,t)\right).$$

This corresponds to the case of strong collisions: each collision restores the equilibrium configuration [**35**]. On the other hand, if $\gamma = 1$ then (4.2) corresponds to a Dirac delta, and (4.1) becomes $\partial_t f = 0$, that is, no relaxation occurs [**117**].

One can verify that $\frac{1}{\Gamma}\int_{\mathbb{R}^d} A(v,w)\,f^{eq}(x,v)\,dv = f^{eq}(x,w)$, which means that scattering does not change the equilibrium configuration. At any instant of time t, we also have the following change in average velocity due to collision

$$\langle w\rangle_{after} = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} w\,A(v,w)\,f(v,t)\,dv\,dw = \gamma\int_{\mathbb{R}^d} v\,f(v,t)\,dv = \gamma\,\langle w\rangle_{before}.$$

This result very well explains that with γ close to 1, weak collisions are modelled, and in this regime the Fokker-Planck equation can be derived from the KS model by means of the Kramers-Moyal expansion; see [**86**]. This is the regime that we consider in this chapter.

We remark that the case of negative γ corresponds to the tendency of reversal of the momentum upon collision, and γ close to -1 is appropriate to model reorientation of molecules in liquid water [**73**], and with $\gamma = -1$ the resulting KS models is closely related to the master equation for quantum transport with stochastic telegraph noise [**88**]. For further discussion and additional references see, e.g., [**72**].

Our focus is the solution of a KS optimality system where a Monte Carlo (MC) scheme is used to solve the KS model and the corresponding KS adjoint problem in order to determine a sufficiently accurate gradient that is needed in a gradient-based optimization procedure. This development work requires the implementation of the different components of the optimality system in the statistical MC framework, and this requirement leads to the reformulation of the KS adjoint problem that we propose in this chapter and illustrate in detail.

In the next section, we formulate our KS optimal control problem in the framework of ensemble control problems as in [**12, 32, 33, 34**]. The corresponding first-order optimality conditions are discussed in Section 4.1.2. In this section, we show how the structure of the KS adjoint equation differs from that of a kinetic model, and discuss a reformulation and an approximation of the adjoint KS master equation by a kinetic model that can be solved with a MC scheme. In Section 4.1.3, we illustrate our numerical setting and give a detailed account of our implementation of the MC scheme for the KS model and its adjoint, and discuss our optimization method. For clarity, all components of our solution strategy are also summarized in pseudo algorithms that closely resemble our simulation and optimization code. In Section 4.1.4, we present results of numerical experiments to validate the ability of our framework to construct control functions that drive the evolution of the ensemble of particles to follow a given trajectory and to attain a final configuration. We also give a comparison with a deterministic approach in the one-dimensional case.

4.1.1. Formulation of a Keilson-Storer optimal control problem

In this section, we discuss the formulation of an optimal control problem governed by the KS master equation. This is a modelling process that requires the definition of the control mechanism and the purpose and cost of the control by means of a cost functional. We consider our control problem defined in the time horizon $[0, T]$ and in a two-dimensional velocity space.

We assume a control mechanism in the KS kernel as follows

$$A(v, w; u) = A_0 \exp\left(-\beta\,|w - \gamma v + u|^2\right), \tag{4.3}$$

where the control $u : [0, T] \to \mathbb{R}^2$ is a time-dependent function representing a velocity field acting on the particles. Clearly, if $u = -(1-\gamma)\mu$ is a constant velocity vector, then the setting above results in a shift of the mean velocity to a desired value $\mu \in \mathbb{R}^2$. As mentioned in the introduction of this chapter, this control mechanism could correspond to a change in temperature gradient of the background fluid.

Notice that, as in [**86**], it holds that $\int_{\mathbb{R}^2} A(v, w)\, dw = \Gamma$. Hence, to be consistent with [**86**], we write our controlled KS master equation as follows

$$\partial_t f(v, t) \;=\; C_u[f](v, t), \tag{4.4}$$

where the controlled collision term is given by

$$C_u[f](v, t) = \int_{\mathbb{R}^2} A(w, v; u)\, f(w, t)\, dw \;-\; f(v, t) \int_{\mathbb{R}^2} A(v, w; u)\, dw. \tag{4.5}$$

Assuming a given initial distribution $f_0 \in H^m_k(\mathbb{R}^2)$ with $(m,k) \in \mathbb{N}^2$, $f_0 \geq 0$, and $u \in U$, where

$$U := H^1_T$$

is the control space, we can write the resulting Cauchy problem as follows

$$f(v,t) = f_0(v) + \int_0^t C_u[f](v,s)\,ds. \tag{4.6}$$

Notice that $C_u[\cdot]$ is linear and bounded with respect to the L^∞ space and therefore continuous. Moreover, it defines a compact Schatten-von-Neumann-integral operator; see [**60**].

With these properties, one can prove existence of a non-negative unique solution $f^* \in C_{T^*}\left(H^m_k(\mathbb{R}^2)\right)$ to our Cauchy problem in a time interval $[0,T^*]$, $0 < T^* \leq T$. See also the energy estimate (4.36) and Remark 4.6 below.

Lemma 4.1 (Conservation). *The controlled Keilson-Storer term* (4.5) *is mass conserving. Specifically, it holds for all times* $t \in [0,T]$

$$\int_{\mathbb{R}^d} C_u[f](v,t)\,dv = 0$$

Proof. Since all integrals in the following converge as the solution of the KS Master equation is in $C_T(H^m_k(\mathbb{R}^d))$, it holds that

$$\begin{aligned}
&\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} f(x,w,t)A(w,v)\,dw - \int_{\mathbb{R}^d} f(x,v,t)A(v,w)\,dw\right)dv\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x,w,t)A(w,v)\,dw\,dv \;-\; \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x,v,t)A(v,w)\,dw\,dv = 0,
\end{aligned}$$

where we interchanged the variable names in the last term. □

As a direct consequence, we have that the KS Master equation preserves mass. Since the Master equation provides positivity of the solution [**129**, page 106] , it holds that

$$\begin{aligned}
\int_{\mathbb{R}^d} |f(v,t)|\,dv &= \int_{\mathbb{R}^d}\left(f_0(v) + \int_0^t C_u[f](v,s)\,ds\right)dv\\
&= \int_{\mathbb{R}^d} f_0(v)\,dv + \int_0^t\int_{\mathbb{R}^d} C_u[f](v,s)\,dv\,ds.
\end{aligned}$$

Now, we can apply Lemma 4.1 to obtain that the last term vanishes. Hence, we get

$$\|f(\cdot,t)\|_{L^1(\mathbb{R}^d)} = \|f_0\|_{L^1(\mathbb{R}^d)} \qquad\qquad \forall t \in [0,T^*].$$

Let us turn again to the case $d = 2$. We assume that $T = T^*$ and remark that, for a fixed initial condition f_0, the solution to (4.4)–(4.5) for a given u defines a control-to-state map

$$G : U \to C^1_T(L^2(\mathbb{R}^2)), \qquad\qquad u \mapsto f = G(u),$$

that is well defined and continuous. Notice that our control is not in a parametrized form. In fact, since our control space is infinite dimensional, we are dealing with an infinite dimensional optimization problem.

Next, we discuss the following cost functional that models the objective of the control and its cost

$$J(f,u) = \int_0^T \int_{\mathbb{R}^2} \theta(v,t)\, f(v,t)\, dv\, dt + \int_{\mathbb{R}^2} \varphi(v) f(v,T)\, dv + \frac{\nu}{2} \|u\|_U^2 . \tag{4.7}$$

This functional results from the notion of ensemble and minimum attention control proposed in [**32, 33, 34**]. For the former, the density of the ensemble of trajectories is involved, for the latter a H_T^1 cost aims at controls that are slow-varying in time, which would make them easier to implement in a physical device.

Recall, that we call the first term in (4.7) the tracking term, the second term represents the final observation, and the last term denotes the cost of the control with a control weight $\nu > 0$. The tracking term includes a function θ that represents an attracting potential for the velocities of the particles. Specifically, let us denote with $\eta_D(t)$ a (time-dependent) desired mean velocity profile for the ensemble of our particles. Then, we may choose $\theta(v,t) = \Theta\left(|v - \eta_D(t)|\right)$ such that the global minimum of the tracking part is achieved when all particles have velocity η_D. Similarly, the final observation term can be defined as $\varphi(v) = \Phi\left(|v - \eta_T|\right)$, which corresponds to the requirement that the mean velocity of the particles at final time is close to η_T. In general, we require that θ and φ are bounded from below and locally convex in the neighbourhood of η_D and η_T, respectively. We consolidate these requirements in the following assumption:

Assumption 4.1. *We take $\theta \in L_T^1(L^2(\mathbb{R}^2))$ and $\varphi \in L^2(\mathbb{R}^2)$, bounded from below and attracting, in the sense that the negative gradients of $\varphi(\cdot)$ and $\theta(\cdot,t)$ are pointing to the unique global minimum of $\varphi(\cdot)$, respectively $\theta(\cdot,t)$, for all $t \in [0,T]$.*

Remark 4.1. *Notice that Assumption 4.1 is a global one. However, in fact it is enough to require that θ and φ are locally convex in a neighbour hood of η_D and η_T instead of being attractive.*

The cost of the control should guarantee bounded controls that continuously change in time. For this purpose, we choose

$$\|u\|_U^2 = \|u\|_{H_T^1}^2 = \int_0^T |u(t)|^2\, dt + \int_0^T |u'(t)|^2\, dt,$$

where $u' = \frac{\mathrm{d}}{\mathrm{d}t}u$. This norm guarantees that u is bounded and, by Sobolev embedding, continuous. The value of the weight $\nu > 0$ establishes the relative importance of achieving the given target with respect to the cost of the corresponding control.

Notice that increasing the value of ν gives more importance to the cost of the control rather than on its tracking ability.

Now, we can formulate our KS optimal control problem

$$
\begin{aligned}
&\min J(f,u) := \int_0^T \int_{\mathbb{R}^2} \theta(v,t)\, f(v,t)\, dv\, dt + \int_{\mathbb{R}^2} \varphi(v) f(v,T)\, dv + \frac{\nu}{2}\, \|u\|^2_{H^1_T} \\
&\text{s.t.} \begin{cases} \partial_t f(v,t) = C_u[f](v,t), & \text{in } \mathbb{R}^2 \times (0,T] \\ f(v,0) = f_0(v) & \text{in } \mathbb{R}^2 \end{cases} \\
&u \in U.
\end{aligned} \tag{4.8}
$$

We remark that, by means of the control-to-state map, this constrained optimization problem can be reformulated as the following unconstrained minimization problem

$$\min_{u \in U} J_r(u), \tag{4.9}$$

where $J_r(u) = J(G(u), u)$ defines the so-called reduced cost functional.

In this setting, we can state existence of a minimizer, which is an optimal control, to (4.9) as follows.

Theorem 4.1. *Let θ and φ fulfil Assumption 4.1 and $f_0 \in H^m_k(\mathbb{R}^2)$, $(m,k) \in \mathbb{N}^2$. Then the KS optimal control problem* (4.8) *has at least one solution $u^* \in U$ with corresponding optimal state $f^* \in C_T(H^m_k(\mathbb{R}^2))$.*

Proof. The functional J given in (4.7) is well-defined for $(f,u) \in C_T(H^k_m(\mathbb{R}^2)) \times H^1_T$. It is bounded from below and coercive in u. Since the map G takes its values in a bounded set of $L^\infty_T(H^k_m(\mathbb{R}^2))$, it follows that J_r is bounded. Furthermore, J_r is weakly lower semi-continuous since G is continuous, the last term is a norm, so it is weakly lower semi-continuous, and the first two terms are linear in f, and then they are weakly continuous with respect to the $L^\infty_T(L^2)$ and L^2 topologies. Thus, we have that, if $\big(u_n\big)_n \subset U$ is a sequence which converges weakly to a $u \in U$, we have

$$\liminf_{n\to+\infty} J_r(u_n) \;=\; \liminf_{n\to+\infty} J\Big(G(u_n), u_n\Big) \;\geq\; J\Big(G(u), u\Big) \;=\; J_r(u)\,.$$

At this point, proving the existence of a minimizer for J_r is standard: Let us take a minimizing sequence $\big(u_n\big)_n \subset U$. This sequence is bounded by the coerciveness of the cost functional in u. Since U is a Hilbert space, we can extract a weakly convergent subsequence, which we do not relabel for simplicity; let us call $u^* \in U$ its limit-point. Then, by the weak-lower semi-continuity of J_r, we can conclude that u^* is a minimizer. □

4.1.2. Keilson-Storer optimality system

In this section, we discuss the optimality system for (4.8) using the Lagrange framework. Since (4.8) and (4.9) are equivalent, assuming Fréchet differentiability of G and J, the first-order necessary condition for a solution to (4.9) is given by

$$\nabla_u J_r(u) = 0,$$

where ∇_u denotes the gradient with respect to u. This fact allows the characterization of an optimal control and makes it possible to develop gradient-based schemes for its computation.

The computation of $\nabla_u J_r(u)$, that is, of $\nabla_u J(G(u), u)$ involves the Fréchet differentiability of G and J, which we assume in this chapter; however, see the previous chapters.

Remark 4.2. *Notice that to transfer the results of the previous chapters, one needs to carefully deal with the regularity of the control that is now depending on x. To be able to apply the last chapters, the drift should fulfill the corresponding assumptions* (2.11) *and therefore the control should be at least one time differentiable. This can be achieved using large enough m and k (depending on the dimension d) and certain Sobolev embeddings.*

A convenient way to obtain $\nabla_u J_r(u)$ is to introduce the Lagrange function corresponding to (4.8) as follows

$$\mathcal{L}(f, u, q, q_0) := J(f, u) + \int_0^T \int_{\mathbb{R}^2} \Big(\partial_t f(v, t) - C_u[f](v, t)\Big)\, q(v, t)\, dv\, dt \\ + \int_{\mathbb{R}^2} \Big(f(v, 0) - f_0(v)\Big)\, q_0(v)\, dv,$$

where q and q_0 represent Lagrange multipliers. In this framework, the so-called optimality system is obtained by requiring that the Fréchet derivatives of L with respect to each of its arguments are zero. This optimality system consists of three parts: 1) the KS model with its initial condition; 2) an adjoint KS master equation that evolves backward in time starting from a terminal condition and having a non-homogeneity related to the cost functional and a linear reaction term related to the collision kernel; 3) the optimality condition $\nabla_u J_r(u) = 0$, which is expressed in terms of the solutions to the two equations discussed in 1) and 2).

We obtain the following adjoint KS (AKS) kinetic model

$$
\begin{aligned}
-\partial_t q(v,t) &= \int_{\mathbb{R}^2} A(v,w;u)\, q(w,t)\, dw \\
&\quad -q(v,t) \int_{\mathbb{R}^2} A(v,w;u)\, dw - \theta(v,t), \qquad \text{in } \mathbb{R}^2 \times [0,T) \\
q(v,T) &= -\varphi(v), \qquad \text{in } \mathbb{R}^2.
\end{aligned}
\tag{4.10}
$$

Notice that the minus sign in front of the time derivative and the terminal condition at $t = T$ result by integration by parts required in the process of derivation of this equation. Making the transformation $s = T - t$, the sign mentioned above is reversed and the equation evolves forward in the new time variable s. This fact is used in the numerical integration of the AKS problem.

However, we are confronted with the problem that (4.10) does not have the structure of a kinetic equation, that is, the right-hand side of the the AKS equation cannot be interpreted as a gain-loss term. The required structure can be partially recovered with the following procedure.

We define

$$
C_0^* = \int_{\mathbb{R}^2} \Big(A(w,v;u) - A(v,w;u) \Big)\, dw,
\tag{4.11}
$$

and, by transformation of the integration variable, we obtain that $C_0^* = \Gamma\, \frac{1-\gamma}{\gamma}$. Further, we define

$$
C_u^*[q](v,t) = \int_{\mathbb{R}^2} A^*(w,v;u)\, q(w,t)\, dw - q(v,t) \int_{\mathbb{R}^2} A^*(v,w;u)\, dw,
\tag{4.12}
$$

where $A^*(w,v;u) = \frac{1}{\gamma}\, A(v,w;u)$, which gives an 'adjoint' mean free time $\tau_q = \gamma\, \tau$. Then, we can write the adjoint KS equation as follows

$$
-\partial_t q(v,t) = C_u^*[q](v,t) + C_0^*\, q(v,t) - \theta(v,t).
\tag{4.13}
$$

Therefore, we need to augment the MC simulation procedure to accommodate the presence of reaction and source terms in (4.13); see the discussion below.

To conclude the formulation of the optimality system, we illustrate the optimality condition. For this purpose, we compute the derivative of the KS collision term with respect to u. In our two-dimensional setting, we have

$$
\begin{aligned}
\partial_{u_\iota} C_u[f](v,t) &= \int_{\mathbb{R}^2} f(w,t)\, A(w,v;u)\, (-2\beta(v_\iota - \gamma w_\iota + u_\iota))\, dw \\
&\quad - f(v,t) \int_{\mathbb{R}^2} A(v,w;u)\, (-2\beta(w_\iota - \gamma v_\iota + u_\iota))\, dw, \qquad \iota = 1,2.
\end{aligned}
\tag{4.14}
$$

Now, we can write the optimality condition given by the gradient of the Lagrange function with respect to u set equal to zero. We have

$$\nu\, u(t) - \nu\, u''(t) - \int_{\mathbb{R}^2} q(v,t)\, \partial_u C_u[f](v,t)\, dv = 0, \tag{4.15}$$

where $u'' = \frac{d^2}{dt^2}u$. This equation should be understood in the weak sense, and for its solution initial and terminal conditions for u are required. Choosing homogeneous Dirichlet boundary conditions corresponds to the case where the control is switched on at $t = 0$, and switched off at $t = T$. Notice that the left-hand side of (4.15) represents the reduced gradient in the $L^2_T(\mathbb{R}^2)$ space. We discuss the construction of $\nabla_u J_r(u)$ in Section 4.2.5.

Summarizing, the KS optimality system is given by

$$\begin{aligned} &\partial_t f(v,t) = C_u[f](v,t), && f(v,0) = f_0(v) \\ &-\partial_t q(v,t) = C_u^*[q](v,t) + C_0^*\, q(v,t) - \theta(v,t), && q(v,T) = -\varphi(v) \\ &-u''(t) + u(t) = \frac{1}{\nu} \int_{\mathbb{R}^2} q(v,t)\, \partial_u C_u[f](v,t)\, dv, && u(0)=0,\ u(T)=0. \end{aligned} \tag{4.16}$$

Our aim is to solve this system using a MC methodology and a gradient-based optimization scheme. For the former task, we investigate a kinetic reformulation of the AKS equation. For this reason, we interpret q as a distribution function and, correspondingly, we refer to adjoint particles. We remark that the main purpose of the adjoint variable q is to allow the computation of the reduced gradient. Therefore, also a suitable approximation of q that results in an effective gradient is appropriate for our goal.

The main differences between the KS and AKS equations are the presence of a linear reaction term $C_0^* q(v,t)$ and of a source term $\theta(v,t)$ in the latter equation. For this reason, we propose the following augmentation of the MC framework.

We have already mentioned that we can reverse the time in the evolution modelled by the AKS equation with the transformation $s = T - t$. Thus, the AKS equation takes the form

$$\partial_s q(v,s) = C_u^*[q](v,s) + C_0^* q(v,s) - \theta(v,s), \qquad q(v,0) = -\varphi(v). \tag{4.17}$$

Next, to accommodate the source term, we choose it given by a negative Gaussian distribution, which is consistent with our Assumption 4.1, as follows

$$\theta(v,s) = -\frac{C_\theta}{2\pi\sigma_\theta^2} \exp\left(-\frac{|v - \eta_D(s)|^2}{2\sigma_\theta^2} \right), \qquad C_\theta > 0, \quad \sigma_\theta > 0,$$

where C_θ represents a weight of the tracking part of the cost functional, and σ_θ has the significance of a standard deviation. This choice is motivated by the fact that we can implement this term in a MC framework by adding adjoint particles to the distribution q in every time-step based on the given Gaussian distribution. Thus, at each time-step, we add a certain number of particles N_{frac} obeying $\mathcal{N}_2(\eta_D(s), \sigma_\theta^2 I_2)$, where I_2 is the two-dimensional identity matrix, $\mathcal{N}_2(\mu, \Sigma)$ is the bivariate normal Gaussian distribution with mean $\mu \in \mathbb{R}^2$ and covariance matrix $\Sigma \in \mathbb{R}^{2\times 2}$.

Similarly, we choose

$$\varphi(v) = -\frac{C_\varphi}{2\pi\sigma_\varphi^2} \exp\left(-\frac{|v-\eta_T|^2}{2\sigma_\varphi^2}\right), \qquad C_\varphi > 0, \qquad \sigma_\varphi > 0.$$

Now, concerning the linear reaction term, we have that

$$\partial_s q(s, v) = C_0^* \, q(v, s),$$

which can be approximated by Euler's method by

$$q(v, s + \delta s) = q(v, s) + \delta s \, C_0^* \, q(v, s),$$

for small $\delta s > 0$. Therefore, we can implement the contribution of the linear reaction term as an increase of the distribution of the adjoint particles at every point according to the factor $(1 + \delta s \, C_0^*)$ for each time-step δs. This means that one particle with velocity v at time s is replaced by $(1 + \delta s \, C_0^*)$ particles with the same velocity v at the time $s + \delta s$. Notice that in our numerical experiments we have $(1 + \delta s \, C_0^*) \approx 2$. It is clear that the value of C_0^* results from the specification of the KS kernel. On the other hand, the time-step δs is chosen in relation to the collision frequency.

4.1.3. A Monte Carlo scheme and numerical optimization

In this section, we illustrate our implementation of the Monte Carlo scheme for our KS problem; see [**70**] for more details. We remark that this is a mesh-less scheme where the particles are represented by labelled pointers to structures that contain all information as velocity, time of collision, etc.. Essentially, a time-step in a MC procedure consists of changing the content of this structure, e.g., velocity, and adding or subtracting pointers, that is, particles, if required. Now, the content of each structure may change for two reasons. If we have a streaming phenomenon, which is the case of an inhomogeneous model and/or presence of external forces, then position and velocity are changed according to the underlying dynamical system. This is the so-called free flow part. On the other hand, the velocity of each particle may change due to collisions with other similar particles or with much smaller particles that constitute the background, which is the case of KS framework; see our discussion in the Introduction. We remark that this type of dynamics resembles a piecewise-deterministic Markov-process, since for the change of the velocity only the current state is considered, see [**131**].

In this evolution process, the time-step size Δt is chosen to be one order of magnitude bigger than the mean time between two collisions. This is in order to retain the statistical significance of the occurrence of collisions. On the other hand, the time-step size cannot be too large since in this case transient phenomena would be filtered out.

To determine when a particle undergoes a velocity transition, we follow the procedure described in [**85**]. If τ^{-1} is the collision frequency, then $\tau^{-1}dt$ is the probability that a particle has a collision during the time dt. Now, assuming that a particle has a collision at time t, the probability that it will be subject to another collision at time $t+\delta t$ is computed according to a Poisson distribution given by

$$\exp\left(-\int_t^{t+\delta t} \tau^{-1}\, dt'\right) = \exp(-\delta t/\tau).$$

Thus, following a standard approach, and using a uniformly distributed random number r between 0 and 1, one obtains the following

$$\delta t = -\tau \log(r). \tag{4.18}$$

The same formula can be used in the simulation of the AKS evolution by replacing τ with τ_q.

Now, assuming that a collision occurs, we need to determine the new velocity after the collision. Clearly, this output depends on the KS kernel that models collision, which can be written as a multivariate normal distribution as follows

$$\begin{aligned} A(v,w;u) &= \Gamma\frac{\beta}{\pi}\exp\left(-\beta|w-\gamma v+u|^2\right) \\ &= \Gamma\frac{\beta}{\pi}\exp\left(-\frac{1}{2}(w-\gamma v+u)^T 2\beta I_2(w-\gamma v+u)\right) \\ &= \Gamma\frac{1}{2\pi\sqrt{\frac{1}{(2\beta)^2}}}\exp\left(-\frac{1}{2}(w-\gamma v+u)^T 2\beta I_2(w-\gamma v+u)\right) \\ &= \Gamma\mathcal{N}_2\left(\gamma v-u, \frac{1}{2\beta}I_2\right). \end{aligned} \tag{4.19}$$

Similarly, for the MC simulation of the AKS model, we can write the transition probability

$$\begin{aligned}
A^*(v,w;u) &= \frac{1}{\gamma}A(w,v;u) = \frac{\Gamma}{\gamma}\frac{\beta}{\pi}\exp\left(-\beta|v-\gamma w+u|^2\right)\\
&= \frac{\Gamma}{\gamma}\frac{\beta}{\pi}\exp\left(-\beta\gamma^2\left|w-\left(\frac{v}{\gamma}+\frac{u}{\gamma}\right)\right|^2\right)\\
&= \frac{\Gamma}{\gamma}\frac{1}{2\pi\sqrt{\frac{1}{(2\beta\gamma^2)^2}}}\exp\left[-\frac{1}{2}\left(w-\left(\frac{v}{\gamma}+\frac{u}{\gamma}\right)\right)^T 2\beta\gamma^2 I_2\left(w-\left(\frac{v}{\gamma}+\frac{u}{\gamma}\right)\right)\right]\\
&= \Gamma\mathcal{N}_2\left(\frac{v}{\gamma}+\frac{u}{\gamma},\frac{1}{2\beta\gamma^2}I_2\right).
\end{aligned}$$

In the case of the KS model, the velocity of a given particle changes according to the normal distribution (4.19), and since the covariance matrix is diagonal, it is possible to generate the new velocity component-wise according to the corresponding one-dimensional distributions. To achieve this goal, we need to have a standard Gaussian random number generator. If r_1 and r_2 are two independent uniformly distributed random numbers between 0 and 1 then, by means of the Box-Muller formula [**30**], we have that

$$z_1 = \sqrt{-2\log(r_1)}\cos r_2, \qquad z_2 = \sqrt{-2\log(r_1)}\sin r_2,$$

are two random numbers distributed according to $\mathcal{N}(0,1)$. Thus,

$$w_x = \gamma v_x - u_x + z_1/\sqrt{2\beta}, \qquad w_y = \gamma v_y - u_y + z_2/\sqrt{2\beta}$$

are two random numbers distributed according to (4.19). Therefore, (w_x, w_y) represents the new velocity of the particle under consideration with velocity (v_x, v_y) before the collision and subject to the control field (u_x, u_y).

At this point, we can illustrate our MC KS solver with the following algorithm, where the initial condition f_0 is used to initialize the list of particles (pointers) in the sense that it provides the density of the initial distribution of the velocities of the particles. In the initialization, we choose a number of particles N_f. Further, we consider a partition of the time interval $[0,T]$ in N_t subintervals of size $\Delta t = T/N_t$ such that $\Delta t \gg \delta t$. With this setting, we have $t^k = k\Delta t$, for the time of the k-th time-step, $k = 0,\ldots,N_t$.

In our implementation, we define F as the list of labelled pointers to structures that resemble particles. We denote with $F^k[p]$ the pointer to the p-th particle at the k-th time-step. We have $p = 1,\ldots,N_f$ and $k = 0,\ldots,N_t$. Furthermore, let $F^k[p].v$ be the velocity of the p-th particle at the k-th time-step, and let $F^k[p].t'$ be the time that is elapsed for the p-th particle starting from t^k. This quantity is used to determine if the particle will undergo another collision in the current time-step, assuming that

$0 \leq F^k[p].t' < \Delta t$. Analogously, we denote with Q the list of labelled pointers to structures representing adjoint particles.

To initialize F^0 using the distribution f_0, we apply Algorithm 4.1 given below. A similar algorithm applies to initialize Q^{N_t} with the distribution $-\varphi$.

Algorithm 4.1 MC KS initialization

```
Require: $f_0(v)$
for $p = 1$ to $N_f$ do
    Compute $F^0[p].v \sim f_0(v)$
    Set $F^0[p].t' = 0$
end for
```

Our MC KS solver is implemented as presented in Algorithm 4.2.

Algorithm 4.2 MC KS equation

```
Require: $f_0(v)$, $u(t)$
Initialize $N_f$ particles using Algorithm 4.1 and $f_0(v)$
for $k = 0$ to $N_t - 1$ do
    for $p = 1$ to $N_f$ do
        while $F^k[p].t' < \Delta t$ do
            Compute $\delta t$ according to (4.18)
            Determine $F^k[p].v \sim \mathcal{N}_2\left(\gamma v - u(t^k), \frac{1}{2\beta} I_2\right)$
            $F^k[p].t' = F^k[p].t' + \delta t$
        end while
        if $F^k[p].t' > \Delta t$ then
            $F^{k+1}[p].t' = F^k[p].t' \bmod \Delta t$
        end if
    end for
end for
```

Next, we present a similar scheme for solving the AKS problem. In this case, we need to implement the contribution to the evolution of the adjoint particles due to the presence of the source term θ and of the linear reaction term. This implementation is illustrated with the following two Algorithms 4.3 and 4.4.

Algorithm 4.3 Implementation of the source term θ at time t^k

```
Require: $\eta_D(t^k)$, $\sigma_\theta$, $N_{frac}$
Generate $N_{frac}$ new particles with velocity components having the normal distribution with mean $\eta_D(t^k)$ and variance $\sigma_\theta$: $v \sim \mathcal{N}\left(\eta_D(t^k), \sigma_\theta^2\right)$
Add these particles to the existing ones in $Q^k$
```

Algorithm 4.4 Implementation of the linear reaction term at time t^k

```
Require: $Q^k$, $N_q^k$, $\Delta t$
for $p = 1$ to $N_q^k$ do
    Generate $\lfloor \Delta t\, C_0^* \rfloor$ particles with the velocity $Q^k[p].v$
end for
Add these particles to the existing ones in $Q^k$
```

Notice that, since in the implementation of the AKS model we vary the number of adjoint particles depending on the linear reaction term and the source term, we index this number with k and write N_q^k. In Algorithm 4.3, we choose $N_{frac} \ll N_f$.

With these two procedures, we can implement the time evolution of the adjoint variable starting from the terminal condition given by $-\varphi(v)$. This function is used to initialize the list of adjoint particles (pointers) in the sense that it provides the density of the initial distribution of the velocities of these particles.

Algorithm 4.5 MC adjoint KS equation

```
Require: $\theta(v,t)$, $\varphi(v)$, $u(t)$.
Initialize $Q^{N_t}$ with $N_q^{N_t} = N_{frac}$ particles using Algorithm 4.1 and $-\varphi$
for $k = N_t$ to $1$ do
    Use Algorithm 4.3 to implement the source term
    Use Algorithm 4.4 to implement the linear reaction term
    for $p = 1$ to $N_q^k$ do
        while $Q^k[p].t' < \Delta t$ do
            Generate $\delta t$ according to (4.18) using $\tau_q$ instead of $\tau$
            Determine $v \sim \mathcal{N}_2\left(\frac{v}{\gamma} + \frac{u(t^k)}{\gamma}, \frac{1}{2\beta\gamma^2} I_2\right)$
            $Q^k[p].t' = Q^k[p].t' + \delta t$
        end while
        if $Q^k[p].t' > \Delta t$ then
            $Q^{k-1}[p].t' = Q^k[p].t' \bmod \Delta t$
        end if
    end for
end for
```

Although a computational mesh to solve the KS and AKS problems is not required, we need this mesh to evaluate the optimization gradient. For this purpose, we consider a bounded domain of velocities $\Upsilon := [-v_{\max}, v_{\max}]^2 \subset \mathbb{R}^2$, where $v_{\max}$ is a working parameter that represents a maximum value of each component of the velocities of the particles. The setting of this parameter is possible since we choose f_0 as Gaussian function that rapidly decays to zero.

Now, we define a partition of Υ in equally-spaced, non-overlapping square cells with side $\Delta v = 2b/N_v$ where $N_v \geq 2$. On this partition, we consider a cell-centred representation of the velocities as follows

$$\Upsilon_{\Delta v} := \left\{ \quad (v_1^i, v_2^j) \in \Upsilon, \quad i, j = 1, \ldots, N_v \quad \right\},$$

where

$$v_1^i = \left(i - \frac{1}{2}\right) \Delta v - v_{\max}, \qquad v_2^j = \left(j - \frac{1}{2}\right) \Delta v - v_{\max}.$$

On the other hand, we recall that on the time interval $[0, T]$, we have the time-steps $t^k = k\Delta t$, $k = 0, \ldots, N_t$, and define

$$\Gamma_{\Delta t} := \left\{ \quad t^k := k\Delta t \in [0, T], \quad k = 0, \ldots, N_t \quad \right\}.$$

Now, we denoted with f_{ij}^k the occupation number of the cell centred in (v_1^i, v_2^j) in the velocity domain. To construct this function, we count the particles at time-step k that have velocity in the cell centred at $v = (v_1^i, v_2^j)$. Thus, we define

$$f_{ij}^k = \sum_{p=1}^{N_f} \mathbb{1}_{ij} \left(F^k[p].v_1, F^k[p].v_2 \right), \tag{4.20}$$

where $\mathbb{1}_{ij}(\cdot, \cdot)$ denotes the indicator function, specifically $\mathbb{1}_{ij}(v_1, v_2) = 1$ if and only if (v_1, v_2) is located in a cell of $\Upsilon_{\Delta v}$ centred at (v_1^i, v_2^j) and zero otherwise.

It results that, if a particle with velocity v within a cell centred at (v_1^i, v_2^j) is subject to collision and acquires a new velocity v' within a cell centred at (v_1^k, v_2^l), then the value of f_{ij} is reduced by 1 and, on the other hand, the value of f_{kl} is increased by 1. Notice that choosing $v_{\max}$ large enough, the probability that the velocity of a particle exceeds the boundary of Υ after collision is very low but possibly not zero. If this rare event happens, we generate again a new velocity for the particle using the same pre-collision velocity as before.

In the case of the AKS evolution, we have to deal with the possibility that the velocity bound $v_{\max}$ will be exceeded in a different way. Because of the structure of the adjoint collision and the high variances of θ and φ, the occurrence of exceeding the bound of $v_{\max}$ by the post-collision velocity of an adjoint particle is not a rare event. However, notice that the purpose of the adjoint distribution is its contribution to the calculation of the gradient given in (4.15). In this formula, it appears inside the integral multiplied with $\partial_u C_u[f]$, and this latter term is vanishing while approaching the computational boundary. Therefore, we only need to consider the adjoint particles that have velocity inside Υ. Specifically, we calculate

$$q_{ij}^k = \sum_{p=1}^{N_q^k} \mathbb{1}_{ij} \left(Q^k[p].v_1, Q^k[p].v_2 \right). \tag{4.21}$$

Now, we focus on (4.14) and assemble the L^2 optimization gradient (4.15) in the vector $g \in \mathbb{R}^{(N_t-1)} \times \mathbb{R}^{(N_t-1)}$. We denote the numerical approximation to $\partial_u C_u[f](v,t)$ at $v = (v_1^i, v_2^j)$ and $t = t^k$ with G_{ij}^k, where $i,j = 0,\dots,N_v$, $k = 1,\dots,N_t$. We use a rectangular quadrature rule to approximate the integrals in (4.14) and obtain

$$\left(G_{ij}^k\right)_1 = (\Delta v)^2 \Big[\sum_{m,n=0}^{N_v} f_{mn}^k A_{mnij}(-2\beta(v_1^i - \gamma v_1^m + u_1^k)) - f_{ij}^k A_{ijmn}^k(-2\beta(v_1^m - \gamma v_1^i + u_1^k))\Big],$$

$$\left(G_{ij}^k\right)_2 = (\Delta v)^2 \Big[\sum_{m,n=0}^{N_v} f_{mn}^k A_{mnij}(-2\beta(v_2^j - \gamma v_2^n + u_2^k)) - f_{ij}^k A_{ijmn}^k(-2\beta(v_2^n - \gamma v_2^j + u_2^k))\Big],$$

where A_{ijmn}^k is given by

$$A_{ijmn}^k \coloneqq \exp\left(-\beta[(v_1^m - \gamma v_1^i + u_1^k)^2 + (v_2^n - \gamma v_2^j + u_2^k)^2]\right).$$

This formula is obtained by inserting $v = (v_1^i, v_2^j)$ and $w = (v_1^m, v_2^n)$ in $A(v,w;u)$.

We can now construct the discrete version of (4.15) using finite differences. The two components of the gradient are given by

$$g_\ell^k \coloneqq \nu u_\ell^k - \frac{\nu}{\Delta t^2}\left(u_\ell^{k+1} - 2u_\ell^k + u_\ell^{k-1}\right) + (\Delta v)^2 \sum_{i,j=0}^{N_v} q_{ij}^k \left(G_{ij}^k\right)_\ell$$
$$k = 1,\dots,N_t - 1, \qquad \ell = 1,2.$$

Notice that this formula provides the numerical approximation to the L^2 gradient while our control field is required in H_T^1. For this purpose, we present the following reasoning that illustrates how to arrive at the formulation of the H^1 gradient.

Consider a Taylor expansion of the reduced cost functional $J_r(u)$ in the Hilbert space X for small $\epsilon > 0$ and $\delta u \in U$ as follows

$$J_r(u + \epsilon\,\delta u) = J_r(u) + \epsilon\,(\nabla J_r(u), \delta u)_X + \frac{\epsilon^2}{2}\left([\nabla^2 J_r(u)]\delta u, \delta u\right)_X + O(\epsilon^3)$$

The actual gradient depends on the choice of which inner product space we use. If we choose the space $X = L^2(0,T;\mathbb{R}^2)$, we have the inner product $(u,v) = \int_0^T u(t)\cdot v(t)\,dt$ and the gradient is given by

$$\nabla J_r(u)|_{L^2} = \nu\, u(t) - \nu\, u''(t) - \int_{\mathbb{R}^2} q(v,t)\,\partial_u C_u[f](v,t)\,dv. \tag{4.22}$$

In the case of $X = H^1(0,T;\mathbb{R}^2)$, we can determine the H^1 gradient based on the fact that the Taylor series must be identical term-by-term regardless of the choice of X. Therefore, we have

$$(\nabla J_r(u)|_{H^1}, \delta u)_{H^1} = (\nabla J_r(u)|_{L^2}, \delta u)_{L^2}.$$

Using the definition of the H^1 inner product $(u,v)_{H^1} = (u,v)_{L^2} + (u',v')_{L^2}$, we obtain the relation

$$\int_0^T \left(\nabla J_r(u)|_{H^1}(t)\, \delta u(t) + \frac{\mathrm{d}}{\mathrm{d}t} \nabla J_r(u)|_{H^1}(t)\, \delta u'(t) \right) dt = \int_0^T \nabla J_r(u)|_{L^2}(t)\, \delta u(t)\, dt,$$

which must hold for all test functions δu. Integrating by parts the second term in the integral on the left-hand side, with the assumption that the control is zero at $t = 0$ and $t = T$, we obtain the following equation for the H^1 gradient.

$$-\frac{\mathrm{d}^2}{\mathrm{d}t^2}[\nabla J_r(u)|_{H^1}(t)] + [\nabla J_r(u)|_{H^1}(t)] = \nabla J_r(u)|_{L^2}(t), \tag{4.23}$$

with the conditions $J_r(u)|_{H^1}(0) = 0$ and $J_r(u)|_{H^1}(T) = 0$. Notice that this is a vector problem for the two components of the gradient $J_r(u)|_{H^1}(t)$. We approximate this problem by a standard finite difference approximation, which results in a block-tridiagonal system. The solution of this system is obtained efficiently by the Thomas method; see [**58**], Algorithm 4.3.

With this preparation, we can formulate the algorithm that provides the $J_r(u)|_{H^1}(t)$ gradient that is required in our optimization scheme.

Algorithm 4.6 Calculate the gradient $\nabla J_r(u)|_{H^1}(t)$

```
Require: control $u(t)$, $f_0(v)$, $\varphi(v)$, $\theta(v,t)$
1: Solve the KS problem using Algorithm 4.2 with inputs $f_0(v)$, $u(t)$
2: Solve AKS problem using Algorithm 4.5 with inputs $-\varphi(v)$, $\theta(v,t)$, $u(t)$
3: Determine the distributions (4.20) and (4.21)
4: Assemble $\nabla J_r(u)|_{L^2}$ according to (4.22)
5: Compute $\nabla J_r(u)|_{H^1}(t)$ solving (4.23)
```

We remark that, with this algorithm, we can implement many different gradient-based optimization schemes [**28**]. In our case, we choose the non-linear conjugate gradient (NCG) method. This is an iterative method that constructs a minimizing sequence of control functions $(u^\kappa)_\kappa$ as illustrated by the following algorithm.

Algorithm 4.7 NCG scheme

```
Require: $u^0(t)$, $f_0(v)$, $\varphi(v)$, $\theta(v,t)$
$\kappa = 0$, $E > tol$
Compute $h^0 = -\nabla J_r(u^0)|_{H^1}$ using Algorithm 4.6
while $E > tol$ and $\kappa < \kappa_{\max}$ do
    Use a line-search scheme to determine the step-size $\alpha_\kappa$ along $h^\kappa$
    Update control: $u^{\kappa+1} = u^\kappa + \alpha_\kappa \, h^\kappa$
    Compute $d^{\kappa+1} = \nabla J_r(u^{\kappa+1})|_{H^1}$ using Algorithm 4.6
    Compute $\beta_\kappa$
    Set $h^{\kappa+1} = -d^{\kappa+1} + \beta_\kappa \, h^\kappa$
    $E = \|u^{\kappa+1} - u^\kappa\|$
    Set $\kappa = \kappa + 1$
end while
return $(u^\kappa, f^\kappa)$
```

In this algorithm, the tolerance $tol > 0$ and the maximum number of iterations $\kappa_{\max} \in \mathbb{N}$ are used as termination criteria. We use backtracking line-search with the Armijo condition. The factor β_κ is based on the Hager and Zhang formula; see [**28**] for more details and references.

An estimate of the computational cost of one iteration in Algorithm 4.7 can be derived as follows: The cost for solving the KS equation is $\mathcal{O}(N_t \, N_f)$, since we calculate for every particle in every time-step a new velocity, similarly $\mathcal{O}(N_t N_f)$ for solving the AKS equation. Furthermore, we have $\mathcal{O}(N_f)$ operations for assembling the distributions on the reference velocity grid for integration, $\mathcal{O}(N_v \times N_v)$ for the integration required for calculating the L^2 gradient in two-dimensions, $\mathcal{O}(N_t)$ to compute the H^1 gradient using the Thomas algorithm. Therefore, the computational complexity of one optimization iteration is $\mathcal{O}(N_t N_f + N_v^2)$. This estimate is without considering line-search, for which $\mathcal{O}(N_t N_v^2)$ operations are required for calculating the functional and $\mathcal{O}(N_t N_f)$ for solving the KS equation for this purpose.

4.1.4. Numerical experiments

In this section, we perform numerical experiments to validate our KS optimal control framework. We assume that initially the particles are at thermal equilibrium obeying the Maxwell-Boltzmann distribution corresponding to the temperature $T_p > 0$ of a gas with particles of mass $M > 0$. This distribution is given by

$$f_0 = \mathcal{N}_2 \left(0, \frac{k_B T_p}{M} I_2 \right),$$

where k_B is the Boltzmann constant. All quantities are given in SI units if nothing else is specified. For the mean collision frequency, we have [**70**]

$$\Gamma = \frac{1}{\tau} = n_m \pi d_p^2 \langle v_{rs} \rangle,$$

where n_m is the number density of background particles, d_p the effective diameter of the particle and $\langle v_{rs} \rangle$ the average relative speed between particles. By the assumptions of the KS kernel, the collision frequency and therefore $\langle v_{rs} \rangle$ are constant over time and space.

Notice that comparing the equations for the mean collision frequency, the conjunction between the parameter A_0 of the KS kernel and the background particles is evident. Specifically, A_0 is proportional to the density of the background particles.

We consider the optimal control problem (4.8) in two dimensions with the following setting

$$\begin{aligned}
\eta_D(t) &= \left(v_A \sin\left(\frac{2\pi}{T} t\right), \frac{v_{\max}}{2T} t \right)^T, & u^0(t) &:= (0,0)^T \\
\theta(v,t) &= -\frac{C_\theta}{2\pi\sigma_\theta^2} \exp\left(-\frac{|v-\eta_D(t)|^2}{2\sigma_\theta^2} \right), & C_\theta &= 10^{20}, \\
\varphi(v) &= -\frac{C_\varphi}{2\pi\sigma_\varphi^2} \exp\left(-\frac{|v-\eta_D(T)|^2}{2\sigma_\varphi^2} \right), & C_\varphi &= 10^{3},
\end{aligned}$$

with $v_A = 250\ m/s$ and $\sigma_\theta = \sigma_\varphi = 10\sigma = 10\sqrt{1/(2\beta)}$. Notice that these parameters determine the ‘width’ of the potentials, that is, their effective basin of attraction. Further, they appear as the variance of the distribution with which particles are created in every time-step, thus their values should be reasonably small.

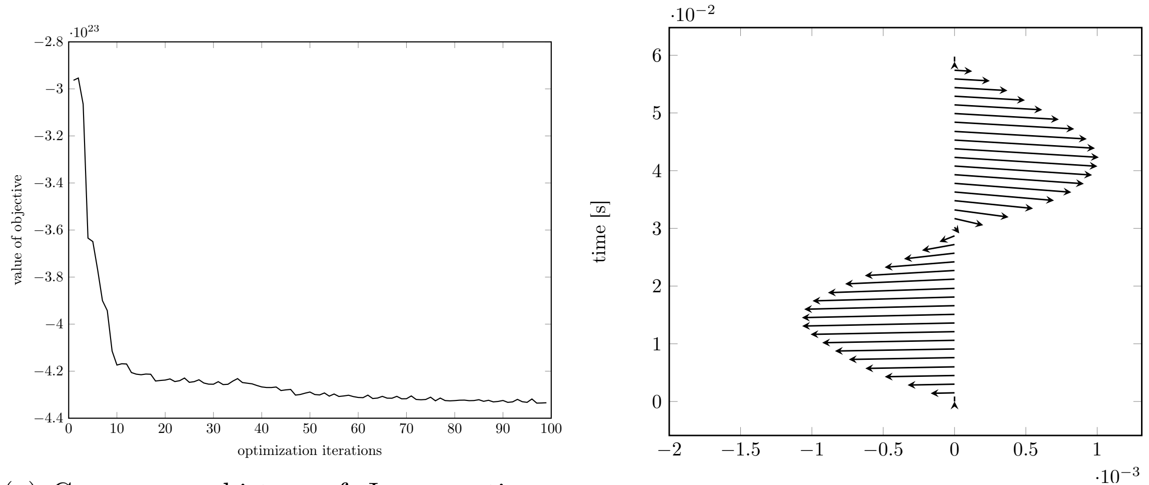
In our simulation, we make the choice of values of the physical and numerical parameters as given in Table 4.1. With the parameter ω_f, we define how many physical particles are combined in a single simulation particle.

After this preparation, we present results of numerical experiments. In Figure 4.1a, we plot the convergence history of the cost functional corresponding to the sequence generated by the NCG scheme. Notice that the randomness in the MC KS solution is responsible for the small fluctuations of the value of J along the minimization process. In Figure 4.1b, we show the resulting optimal control field obtained with the NCG scheme starting from an initial guess $u^0 = 0$. The ordinate axis shows the time and for every discrete point in time, the optimal control vector is plotted. Notice that this vector is zero at the initial and terminal times, consistently with our requirement in the H_0^1 setting.

Next, we plot the action of the optimal control on the evolution of the density f. For this purpose, in Figure 4.2, we plot the value of the mean velocity with time

Symbol	Value	Symbol	Value
T_p $[K]$	700	M $[kg]$	$6.63 \cdot 10^{-26}$
β $[\frac{s^2}{m^2}]$	$\frac{m}{2k_BT}$	k_B $[\frac{m^2\,kg}{s^2\,K}]$	$1.38 \cdot 10^{-23}$
γ [-]	0.9	d $[m]$	$0.4 \cdot 10^{-3}$
ω_f [-]	10	N_f [-]	$5 \cdot 10^4$
N_t [-]	40	Δt $[s]$	$1.5 \cdot 10^{-3}$
N_v [-]	20×20	$v_{\max}$ $[\frac{m}{s}]$	10^3
ν $[\frac{s^2}{m^2}]$	10^{-10}	Δv $[\frac{m}{s}]$	105.26
N_{frac} [-]	$0.1 \cdot N_f$	T $[s]$	$6 \cdot 10^{-2}$

Table 4.1. Physical and numerical parameters.



(a) Convergence history of J over optimzation iterations.

(b) Optimal control field. The ordinate axis shows the time.

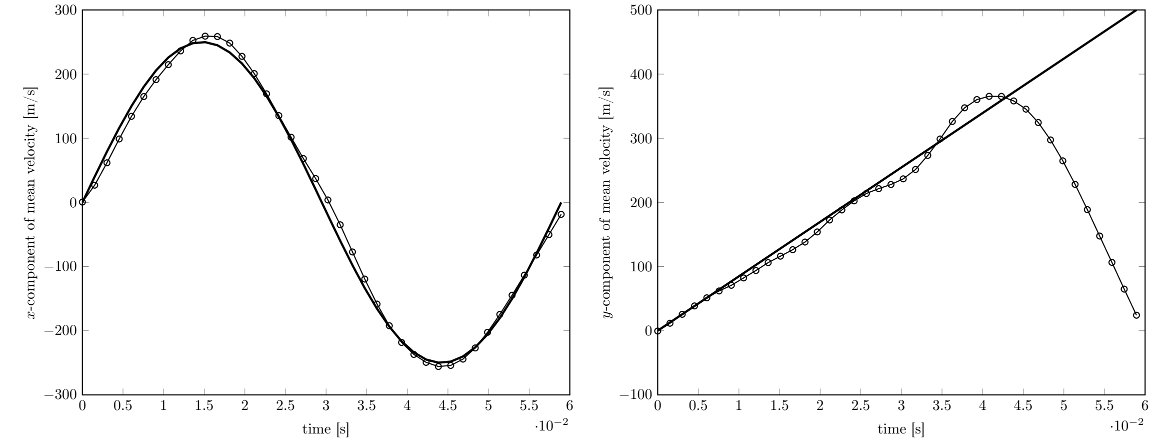
Figure 4.1. Results of numerical experiment in the H^1 case.

computed using the resulting distribution as follows

$$\langle v \rangle (t) = \int v\, f(v,t)\, dv.$$

In Figure 4.2a, we show the x component of the mean velocity and compare it to the desired velocity. Similarly, in Figure 4.2b, we plot the y component of the mean velocity. Notice that, since we require that the control is zero at final time, the y component of the desired velocity is not attainable, which explains the behaviour depicted in Figure 4.2b.

However, our choice of having the control field equal to zero at $t = 0$ and $t = T$ is arbitrary and can be replaced by other conditions to define different control spaces. Moreover, boundary conditions are not required if our control space is chosen as



(a) Comparison between mean velocity (—o—) and desired one (—) (x component).

(b) Comparison between mean velocity (—o—) and desired one (—) (y component).

Figure 4.2. Results of numerical experiment using H^1 control: evolution of the mean velocity. The axis of ordinates shows the velocity and the axis of abscissas shows the time, both in SI units.

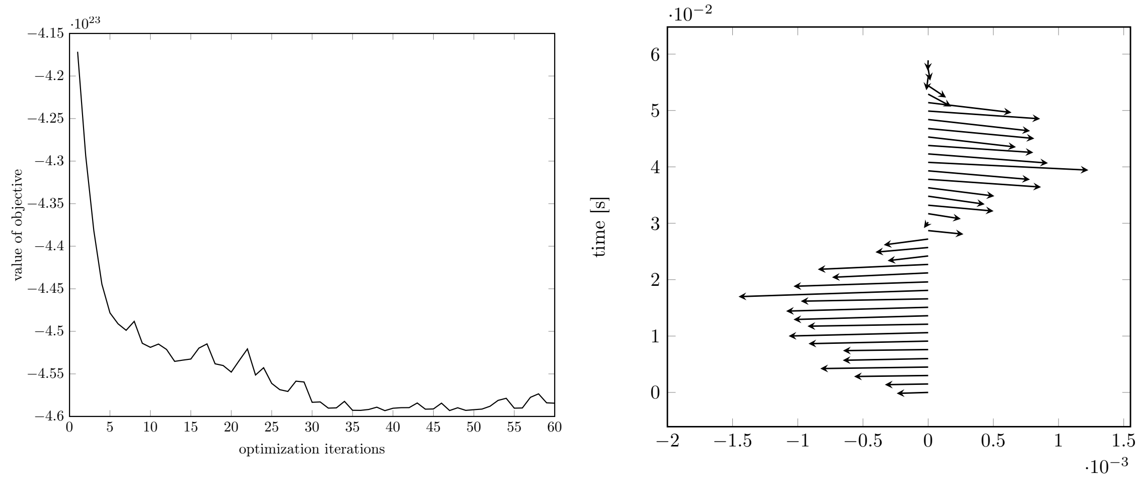
$U = L^2(0,T;\mathbb{R}^2)$. In this case, we use the L^2 gradient given by

$$\nabla J_r(u)|_{L^2} = \nu\, u(t) - \int_{\mathbb{R}^2} q(v,t)\, \partial_u C_u[f](v,t)\, dv,$$

which results in a NCG scheme where $\nabla J_r(u)|_{H^1}$ is replaced by $\nabla J_r(u)|_{L^2}$.

With this version of Algorithm 4.7, we also obtain a sequence of controls that minimizes the objective functional as shown in Figure 4.3a, and determine the optimal controls depicted in Figure 4.3b. Comparing this latter result with that shown in Figure 4.3b, we see that the optimal control is less regular in the L^2 case, as expected. On the other hand, with this setting, the desired velocity profile is attainable and our control framework can track the desired velocity for all times as shown in Figure 4.4.

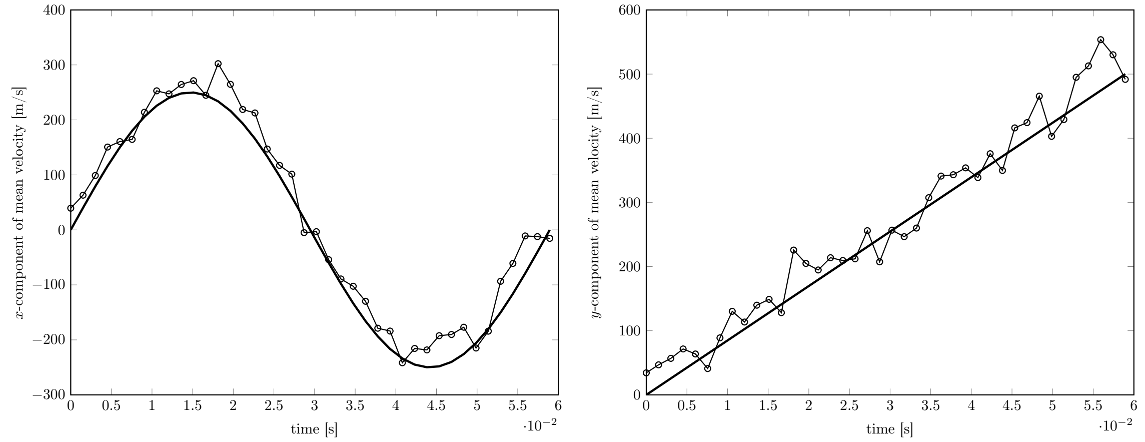
The last part of this section presents a comparison of our MC framework with a similar optimal control approach where finite differences approximation, explicit Euler time-stepping scheme, and second-order quadrature are used. For this purpose, we consider a one-dimensional setting and L^2 control costs. In the deterministic counterpart of our scheme, we consider a uniform mesh in the velocity space. For stability of the forward Euler time-stepping procedure, we need to choose a time-step-size that is one order of magnitude smaller than in the MC method. Notice that the latter is inherently numerically stable [**132**]. In both cases, we use a gradient-descent scheme with fixed step-size for updating the control. As termination criterion, we stop both algorithms when the reduction of the value of the cost functional between two consecutive optimization steps is less than 10^{-4}.



(a) Convergence history of J over optimization iterations in the L^2 case.

(b) Optimal control field in the L^2 case. The ordinate axis shows the time.

Figure 4.3. Results of numerical experiments in the L^2 case.



(a) Comparison between mean velocity (—o—) and desired one (—) (x component).

(b) Comparison between mean velocity (—o—) and desired one (—) (y component).

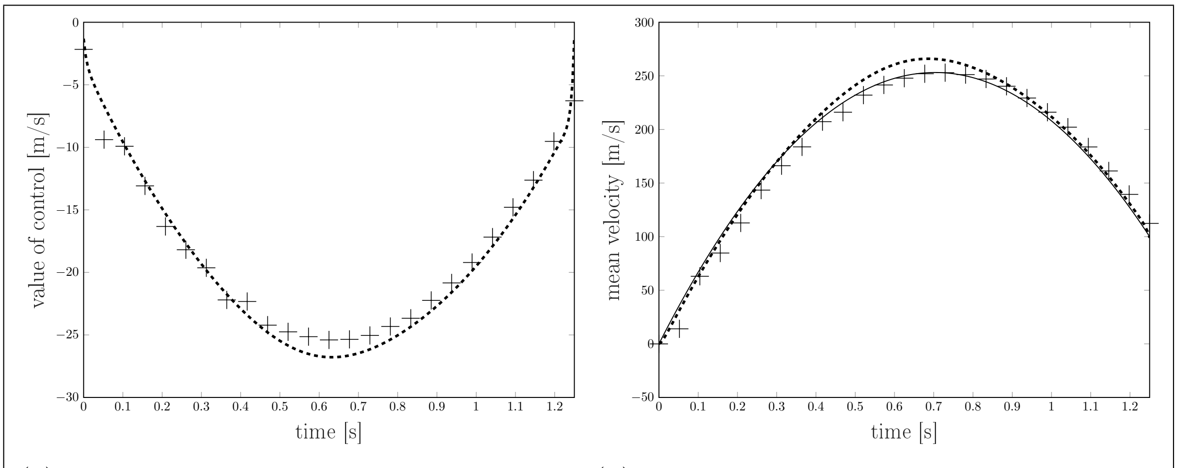
Figure 4.4. Results of numerical experiment using L^2 controls: evolution of the mean velocity. The axis of ordinates shows the velocity and the axis of abscissas shows the time, both in SI units.

In this experiment, we choose the following desired velocity profile

$$\eta_D(t) = v_1 \left(\frac{t}{T}\right)^2 + v_2 \frac{t}{T},$$

where $v_1 = -800$ m/s, $v_2 = 900$ m/s and $T = 1.25$ s.

In Figure 4.5a, we compare the optimal controls obtained with these two methods. We see that they match very well. In Figure 4.5b, we plot the corresponding velocity profiles, compared with $\eta_D(t)$, showing comparable results.



(a) Control calculated with deterministic (- - -) and MC (+++) scheme in the L^2 case.

(b) Resulting trajectory of the mean velocity calculated with deterministic (- - -) and MC (+++) and the desired mean velocity (——).

Figure 4.5. Comparison of results obtained with a deterministic scheme and with the MC scheme. The axis of ordinates shows the value of the control (left) and the mean velocity (right) and the axis of abscissas shows the time, all in SI units.

4.2. Controlled linear kinetic models

It is the purpose of this section to investigate an optimal control problem governed by the following kinetic model

$$\begin{aligned} (4.24)\quad & \partial_t f(x,v,t) + v\cdot\nabla_x f(x,v,t) + u(x)\cdot\nabla_v f(x,v,t) \\ &= \int_{\mathbb{R}^d} A(w,v)\, f(x,w,t)\,dw - f(x,v,t)\int_{\mathbb{R}^d} A(v,w)\,dw. \end{aligned}$$

In this model, we consider the function $u(x):\mathbb{R}^d\to\mathbb{R}^d$ as the control force, and we choose again the representative Keilson-Storer collision term given as the right-hand side of (4.1) . Notice that, as in [**86**], it holds $\int_{\mathbb{R}^d} A(v,w)\,dw = \Gamma$. Hence, to be consistent with [**86**], we omit to write Γ on the right-hand side of (4.24); cf. (4.4). Moreover, we assume $u\in U$ with control space $U\subseteq L^2(\Omega)$.

In Section 4.2.1, we derive the equations for the important moments of the linear kinetic model (4.24). These moments are of importance since they can be interpreted as macroscopic properties of the system of particles they describe. In Section 4.2.2, techniques of Chapter 2 are exploited to gain well-posedness results for linear kinetic models using the Keilson-Storer collision term.

In the Section 4.2.3, we formulate our kinetic optimal control-in-the-force problem in the framework of ensemble control problems [**12, 32**]. In particular, we illustrate the construction of our cost functional with ensemble trajectory- and final-configuration terms and a cost of the control. For this setting, we prove existence of solutions in the given control space.

In Section 4.2.4, we discuss the first-order optimality conditions that must be satisfied by an optimal control. Thus, we derive the corresponding optimality system and analyse the structure of the resulting adjoint equation. Through this analysis, we elaborate on this equation to obtain an equivalent adjoint problem with gain-loss structure, a linear reaction term and a source term. In this process, we identify this source term and the terminal condition for the adjoint problem as constitutive parts of the cost functional. Specifically, we notice that these terms must, on the one hand, play the role of attractive potentials for the ensamble of particles and, on the other hand, have the properties of a probability density function in order to accommodate them in the MC scheme.

In Section 4.2.5, we present our optimization strategy. We give a detailed account of our implementation of the MC scheme for solving (4.24) and our adjoint model in order to assemble the optimization gradient. This gradient is then used to implement a non-linear conjugate gradient scheme; see, e.g., [**28**]. For clarity, all components of our solution strategy are also summarized in pseudo algorithms that closely resemble our simulation and optimization codes. In Section 4.2.6, we present results of numerical experiments and verifications that approves the ability of our framework to construct optimal control forces that drive the evolution of the ensemble of particles to follow a desired trajectory in phase space and to attain a final configuration.

4.2.1. Moments of linear kinetic models

We start with giving details for the calculation of the moments in one dimension in space and velocity of the controlled linear model given in (4.24). Notice that in this case the operator ∇_x can be identified with ∂_x. Further, the investigation in higher dimensions can be performed analogously since the Keilson-Storer collision kernel assumes no correlation between the different dimensions in velocity space.

We assume that f and ∇f approach 0 for $|v| \to \infty$ sufficiently fast, such that the following calculations hold true. In particular, it is sufficient to assume that f_0 is an element in the weighted Sobolev-space $f \in H^1_{n+1}((0,T);\Omega \times \mathbb{R})$ for the $n-$th moment and the drift increases at most linearly.

We define the following moment of the collision kernel (4.2) as

$$\begin{aligned} A^0 &:= \int A(v,w)\,dv = A_0\sqrt{\frac{\pi}{\beta}}, \\ A^1(w) &:= \int v\,A(v,w)\,dv = A_0\sqrt{\frac{\pi}{\beta}}\gamma\,v = A^0\gamma\,w, \qquad (4.25) \\ A^2(w) &:= \int v^2\,A(w,v)\,dv = \frac{A^0}{2\beta} - A^0 w^2\gamma^2. \end{aligned}$$

Notice that in (4.25), the zeroth-moment is in fact the collision frequency and does not depend on the velocity. The moments that are worth to investigate, arise if we multiply (4.24) with collision invariants of the Boltzmann equation and integrate over v afterwards, see [**110**]. These collision invariants are 1, v, $|v|^2$ since for these functions the term corresponding to the collision in the Boltzmann equation vanishes in the moment equation. The resulting moments can be used to calculate important macroscopic properties of the system of particles.

We define the zeroth, first and second moment n_f, $\bar{m}$ and $\bar{\sigma}$ of f as

$$\begin{aligned} n_f(x,t) &:= \int_{\mathbb{R}} f(x,v,t)\,dv, \qquad \bar{m}(x,t) := \int_{\mathbb{R}} v\,f(x,v,t)\,dv, \\ \bar{\sigma}(x,t) &:= \int_{\mathbb{R}} v^2\,f(x,v,t)\,dv \end{aligned} \tag{4.26}$$

and present how to derive evolution equations for them.

We start with the zeroth moment. Integration of (4.24) over v leads to

$$\begin{aligned} &\int_{\mathbb{R}} \partial_t f(x,v,t)\,dv + \int_{\mathbb{R}} v\cdot\nabla_x f(x,v,t)\,dv + \int_{\mathbb{R}} u(x)\cdot\nabla_v f(x,v,t)\,dv \\ &\qquad = \int_{\mathbb{R}} \Big(\int A(w,v)\,f(x,w,t)\,dw \;-\; f(x,v,t) \int A(v,w)\,dw \Big)\,dv. \end{aligned}$$

Since we have choose $d=1$ in this section, the operators ∇_x and ∇_v could be replaced by ∂_x and ∂_v. The third term on the left-hand side of (4.24) vanishes by integration by parts, since u does not depend on v. The term on the right-hand side vanishes by virtue of Lemma 4.1. Therefore, we can write

$$\partial_t n_f(x,t) + \partial_x \bar{m}(x,t) = 0. \tag{4.27}$$

This is precisely what can be calculated for the Boltzmann equation, see [**110**]. Since the constant function 1 is an invariant of the the KS collision term and the Boltzmann collision term, this is what one expects.

Now we turn to the first moment. Applying $\mathbb{E}[v]$ to (4.24) leads to

$$\begin{aligned} &\int_{\mathbb{R}} v\Big(\partial_t f(x,v,t) + v\cdot\nabla_x f(x,v,t) + u(x)\cdot\nabla_v f(x,v,t)\Big)\,dv \\ &\qquad = \int_{\mathbb{R}}\int_{\mathbb{R}} vA(w,v)\,f(w,x,t)\,dw\,dv - \int_{\mathbb{R}} v\,f(x,v,t) \int_{\mathbb{R}} A(v,w)\,dw\,dv. \end{aligned} \tag{4.28}$$

We calculate the integral for each term. It holds for the first term and second term on the left-hand side of (4.28) with interchanging differentiation and integration

$$\begin{aligned} &\int_{\mathbb{R}} v\,\partial_t f(x,v,t)\,dv = \partial_t \int_{\mathbb{R}} v\,f(x,v,t)\,dv = \partial_t \bar{m}(x,t), \\ &\int_{\mathbb{R}} v^2\cdot\nabla_x f(x,v,t)\,dv = \partial_x \int_{\mathbb{R}} v^2\,f(x,v,t)\,dv = \partial_x \bar{\sigma}(x,t) = \partial_x\Big(2n_f(x,t)E_{sp}\Big), \end{aligned}$$

where n_f is the is the number density and E_{sp} is the specific energy, see [**122**].

Further, it holds for the third term on the left-hand side of (4.28), with integration by parts, assuming that the boundary term vanishes since f, ∇f decay sufficiently fast with respect to velocity at infinity

$$\int_{\mathbb{R}} v\ (u(x) \cdot \nabla_v f(x,v,t))\ dv = -u(x) \int_{\mathbb{R}} 1\, f(x,v,t)\, dv = -u(x)\, n_f(x,t),$$

It holds for the first term on the right-hand side of (4.28) with Fubini's theorem and (4.25) that

$$\begin{aligned}\int_{\mathbb{R}}\int_{\mathbb{R}} vA(w,v)\, f(w,x,t)\, dw\, dv &= \int_{\mathbb{R}} f(x,w,t) \int_{\mathbb{R}} v\, A(w,v)\, dv\, dw \\ &= \int_{\mathbb{R}} A^1(w)\, f(x,w,t)\, dw \\ &= A^0 \gamma\, \bar{m}(x,t).\end{aligned}$$

Similarly, it holds for the second term on the right-hand side

$$\int_{\mathbb{R}} v\, f(x,v,t) \int_{\mathbb{R}} A(v,w)\, dw\, dv = A^0 \bar{m}(x,t). \tag{4.29}$$

With the calculations above, we can write down the equation of the evolution of the first momentum $\bar{m}(x,t)$ of (4.24)

$$\partial_t \bar{m}(x,t) + \partial_x \bar{\sigma}(x,t) = u(x) n_f(x,t) - A^0(1-\gamma)\bar{m}(x,t). \tag{4.30}$$

The last part on the right-hand side of the equation is due to the collision term. Notice that without a force, that is $u \equiv 0$, and assuming that the specific energy is uniformly distributed over space (and hence not depending on x), the mean approaches zero eventually, since $A^0(1-\gamma) > 0$ for $\gamma \in (0,1)$. Further, $\bar{m}$ has the physical interpretation of being proportional to the macroscopic velocity.

We finish this section with the derivation of the second moment. For this goal, we apply $\mathbb{E}[v^2]$ to (4.24). Notice that this is not the centralized moment. However, up to a constant it can be interpreted as the energy density of the system. It follows

$$\begin{aligned}&\int_{\mathbb{R}} v^2\, \partial_t f(x,v,t)\, dv + \int_{\mathbb{R}} v^3\, \nabla_x f(x,v,t)\, dv + \int_{\mathbb{R}} u(x)\, v^2\, \nabla_v f(x,v,t)\, dv \\ &\qquad = \int_{\mathbb{R}} v^2 \int_{\mathbb{R}} A(w,v)\, f(x,w,t)\, dw\, dv - \int_{\mathbb{R}} v^2\, f(x,v,t) \int_{\mathbb{R}} A(v,w)\, dw\, dv.\end{aligned}$$

We perform calculations as above and omit to present all the details here. Notice that

$$\int_{\mathbb{R}} v^3\, \nabla_x f(x,v,t)\, dv = \partial_x \int_{\mathbb{R}} v|v|^2 f(x,v,t)\, dv.$$

The integrand on the right-hand side is connected to the heat flux, see [**122, 110**]. We define the quantity $\bar{h}(x,t)$ as

$$\bar{h}(x,t) \coloneqq \int_{\mathbb{R}} v|v|^2 f(x,v,t)\, dv.$$

Specifically, $\bar{h}$ is the heat flux, if in $v|v|^2$, v is replaced by the difference of the v and the macroscopic velocity and can moreover be interpreted as the third non-centralized moment.

Further, by integration by parts,

$$u(x)\int_{\mathbb{R}} v^2\,\nabla_v f(x,v,t)\,dv = -2u(x)\int_{\mathbb{R}} v\,f(x,v,t)\,dv = -2u(x)\,\bar{m}(x,t).$$

For the first term on the right-hand side, we obtain with the moments of the kernel given in (4.25)

$$\begin{aligned}\int_{\mathbb{R}} v^2\int_{\mathbb{R}} A(w,v)\,f(w,v,t)\,dw\,dv &= \int_{\mathbb{R}} f(x,w,t)\int v^2\,A(w,v)\,dv\,dw\\ &= \int_{\mathbb{R}} f(x,w,t)\,A^2(w)\,dw\\ &= \int_{\mathbb{R}} f(x,w,t)\left(\frac{A^0}{2\beta}+A^0\gamma^2w^2\right)\,dw\\ &= \left(\frac{A^0}{2\beta}n_f(x,t)+A^0\gamma^2\bar{\sigma}(x,t)\right).\end{aligned}$$

For the second term it holds that

$$\int_{\mathbb{R}} v^2\,f(x,v,t)\int_{\mathbb{R}} A(w,v)\,dw\,dv = A^0\bar{\sigma}(x,t).$$

Summarizing, we can write an ordinary differential equation for the second moment $\bar{\sigma}$

$$\partial_t\bar{\sigma}(x,t)+\partial_x\bar{h}(x,t) = 2u(x)\bar{m}(x,t)+\frac{A^0}{2\beta}n_f(x,t)-A^0(1-\gamma^2)\bar{\sigma}(x,t). \tag{4.31}$$

Remark 4.3. *From these non-centralized moments, the centralized ones can be calculated. Moreover, the same calculations can be performed components-wise in higher dimensions see,* [**110**]. *Notice that the equation for a moment involves the derivative of a higher moment.*

Remark 4.4. *Moreover, since v, $|v|^2$ are not collision invariants for the Keilson-Storer collision term, we gain additional terms compared to the moments of the Boltzmann equation. This is expected since the velocity and energy of the particles under consideration is influenced by the velocity and energy of the background species via collision and therefore not conserved.*

4.2.2. Well-posedness of linear kinetic models

The aim of this section is to show the well-posedness of (4.24) by deriving energy estimates as in Chapter 2.

We can write a generalized form of (4.24) with the general drift function $a = a(x, v, t)$ as

$$\partial_t f(x,v,t) + a(x,v) \cdot \nabla f(x,v,t) + (\text{div } a(x,v,t))\, f(x,v,t) + \frac{1}{\tau} f(x,v,t) = \int_{\mathbb{R}^d} A(w,v)\, f(x,w,t)\, dw, \tag{4.32}$$

where ∇ is defined as $\nabla = (\nabla_x, \nabla_v)$ and the divergence operator is considered to act on the x and v variable. Moreover, we assume that $a = (a^1(v,t), a^2(x,t))$ with a^1 and a^2 having at most linear growth. Notice that in (4.32) we consider the phase space, and therefore have in fact two variables with which we have to deal fundamentally different. On the one hand, the position-variable x attains only finite values since $x \in \Omega$. On the other hand, the velocity-variable v is unbounded since $v \in \mathbb{R}^d$. For the scope of this section, we supplement (4.32) with absorbing boundary conditions; this taking $f(x,v,t) = 0$ for all $x \in \partial\Omega$.

We can treat the left-hand side of (4.32) analogously to Section 2.2 in Chapter 2. The only change is that the term div a is replaced by div $a + \frac{1}{\tau}$. Recall that $\frac{1}{\tau}$ is the mean collision frequency and given by

$$\frac{1}{\tau} = \int_{\mathbb{R}^d} A(v,w)\, dv.$$

Therefore, $\frac{1}{\tau}$ is constant and only depending on the parameters of collision kernel. Hence, what is left is to investigate the term on the right-hand side of (4.32).

Before we consider the full equation (4.32), we perform estimates for the case of Section 4.1, specifically equation (4.1). Notice that this is taking $a = (v, 0)$ in (4.32) and assuming that f is space-homogeneous for all t, that is $f(x,v,t) = f(v,t)$. Hence, we can rewrite (4.1) as

$$\partial_t f(v,t) \;+\; \frac{1}{\tau} f(v,t) = \int_{\mathbb{R}^d} A(w,v)\, f(w,t)\, dw. \tag{4.33}$$

Taking the inner product of (4.33) with $f(v,t)$ and performing standard computations leads to

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} \|f(t)\|^2_{L^2(\mathbb{R}^d)} - \frac{1}{\tau} \|f(t)\|^2_{L(\mathbb{R}^d)} = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} A(w,v)\, f(w,t)\, f(v,t)\, dw\, dv,$$

where the right-hand side is by means of Hölder's inequality less or equal to

$$\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |A(w,v)|^2\, dv \right)^{1/2} \left(\int_{\mathbb{R}^d} |f(v,t)|^2\, dv \right)^{1/2} f(w,t)\, dw.$$

Therefore, it holds that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} \|f(t)\|^2_{L^2(\mathbb{R}^d)} \le \frac{1}{\tau} \|f(t)\|^2_{L^2(\mathbb{R}^d)} + A_0 \left(\frac{\pi}{2\beta} \right)^{d/4} \|f(t)\|_{L^2(\mathbb{R}^d)} \|f(t)\|_{L^1(\mathbb{R}^d)}\,.$$

Since $f \in C([0,T]; L^2(\mathbb{R}^d))$, an application of Grönwall's lemma leads to the existence of a constant $c > 0$ such that

(4.34)
$$\begin{aligned}\|f(t)\|_{L^2(\mathbb{R}^d)} &\leq c \left(\|f_0\|_{L^2(\mathbb{R}^d)} + A_0 \left(\frac{\pi}{2\beta}\right)^{d/4} \int_0^t \|f(s)\|_{L^1(\mathbb{R}^d)}\, ds \right) \exp\left(\int_0^t \frac{1}{\tau}\, ds\right) \\ &\leq c e^{T/\tau} \left(\|f_0\|_{L^2(\mathbb{R}^d)} + A_0 \left(\frac{\pi}{2\beta}\right)^{d/4} t \right),\end{aligned}$$

where in the last inequality we used the conservation of mass and the non-negativity of f that lead to

$$\|f(t)\|_{L^1(\mathbb{R}^d)} = \|f_0\|_{L^1(\mathbb{R}^d)} = 1, \qquad\qquad t \in [0,T].$$

Remark 4.5. *Notice that the same computations hold true if we exchange* (4.1) *with* (4.4) *since the additional shift in the kernel of A does not change the integral. More in detail, it holds that*

$$\int_{\mathbb{R}^d} A(w,v;u)\, dv = \int_{\mathbb{R}^d} A(w,v)\, dv = \left(\frac{\pi}{\beta}\right)^{d/2}.$$

Now, we turn to energy estimates in H^1_k spaces of the equation (4.33) Taking the derivative with respect to v_l, $l = 1, \dots, d$, and multiplying afterwards with $|v|^k$ gives the equation with

$$\partial_t \left(|v|^k \partial_{v_l} f(v,t)\right) - |v|^k \frac{1}{\tau} \partial_{v_l} f_k(v,t) = \int_{\mathbb{R}^d} f(w,t) |v|^k\, \partial_{v_l} A(w,v)\, dw.$$

Taking the inner-product of this equation with $\partial_{v_l} f(v,t)\, |v|^k$ $k, l \in \mathbb{N}$, and following the same procedure as above, we can estimate the right-hand side as

$$\begin{aligned}\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \left| |v|^k \partial_{v_l} A(w,v) \right|^2 dv \right)^{1/2} \left(\int_{\mathbb{R}^d} \left| |v|^k \partial_{v_l} f(v,t) \right|^2 dv \right)^{1/2} f(w,t)\, dw \\ = \|f(t)\|_{L^2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \lambda^k_l(w)\, f(w,t)\, dw.\end{aligned}$$

In the last equality, $\lambda^k_l(w)$ denotes a polynomial which is for all l of order k at most and defined as

$$\lambda^k_l(w) := \left(\int_{\mathbb{R}^d} \left| |v|^k \partial_{v_l} A(w,v) \right|^2 dv \right)^{1/2}. \tag{4.35}$$

Therefore, it is determined by the parameters of A which are β and γ. Notice that $\partial_{v_l} A(w,v) = -2\beta(v_l - \gamma w_l) e^{-\beta|v-\gamma w|^2}$ and further that the quantity $\int_{\mathbb{R}^d} \lambda^k_l(w)\, f(w,t)\, dw$ can be estimated using the moments of f. We can conclude by summing up over $l = 1, \dots, d$ that

$$\|f(t)\|_{H^1_k(\mathbb{R}^d)} \leq c e^{T/\tau} \left(\|f_0\|_{H^1_k(\mathbb{R}^d)} + \sum_{\kappa=0}^{k} \int_0^t \|f(t)\|_{L^1_\kappa(\mathbb{R}^d)}\, ds \right), \tag{4.36}$$

where $L^1_k(\mathbb{R}^d)$ is defined analogously to $L^2_k(\mathbb{R}^d)$ in Definition 2.1.

REMARK 4.6. *This procedure can also be executed for $H^m_k(\mathbb{R}^d)$ spaces. Notice that also in this case the polynomials $\lambda^k_{l,m}(w)$ are of order at most k, but differ in their coefficients, and are defined as*

$$\lambda^k_{l,m}(w) := \left(\int_{\mathbb{R}^d} \left||v|^k \partial^m_{v_l} A(w,v)\right|^2 \, dv\right)^{1/2}. \tag{4.37}$$

After these preliminary results, we consider again (4.32). Recall, that the Keilson-Storer collision term is equivalent to the BGK model for $\gamma = 0$; see Section 4.1. Hence, (4.32) takes the form

$$\partial_t f(x,v,t) + a(x,v,t)\cdot\nabla f(x,v,t) + \left(\text{div } a(x,v,t)\right) f(x,v,t) + \frac{1}{\tau} f(x,v,t) = \frac{1}{\tau} f^{eq}(x,v),$$

where f^{eq} is a given Maxwell-Boltzmann distribution in velocity and a smooth prescribed distribution in space and independent of time since it is assumed to be a steady state. Notice that f^{eq} is rapidly decaying for $|v| \to \infty$ and bounded with respect to x since we are in a bounded domain Ω. Therefore, we can directly apply the theorems of Chapter 2. In particular, by virtue of Theorem 2.2, we can conclude that

$$\begin{aligned}\|f(t)\|_{H^m(\Omega\times\mathbb{R}^d)} \le C\left(\|f_0\|_{H^m(\Omega\times\mathbb{R}^d)} + \frac{1}{\tau}\int_0^t \|f^{eq}(s)\|_{H^m(\Omega\times\mathbb{R}^d)}\, ds\right) \times \\ \times \exp\left(C\int_0^t \|\nabla a(s)\|_{C^m_b(\Omega\times\mathbb{R}^d)} + \frac{1}{\tau}\, ds\right).\end{aligned}$$

Now, we consider again the KS kernel with arbitrary $\gamma \in (-1,1)$.

As in (2.19), we consider the L^2 scalar product in $\Omega\times\mathbb{R}^d$ of the right-hand side of (4.33) with f and perform the following estimates, during which we apply Young's convolution inequality. For this notice that the KS collision term can be written as a convolution after a variable change $\tilde{w} = \gamma w$ as

$$A(w,v) = \tilde{A}(v-\tilde{w}) := \int_{\mathbb{R}^d} \exp\left(-\beta|v-\tilde{w}|^2\right)\, dv.$$

Then it holds that

$$\begin{aligned}&\iint_{\Omega\times\mathbb{R}^d}\int_{\mathbb{R}^d} A(w,v)\, f(x,w,t)\, dw\, f(x,v,t)\, dv\, dx \\ &\quad\le \frac{1}{\gamma}\int_\Omega \iint_{\mathbb{R}^d\times\mathbb{R}^d} \tilde{A}(v-\tilde{w})\, \tilde{f}(x,\tilde{w},t)\, f(x,v,t)\, d\tilde{w}\, dv\, dx \\ &\quad\le \frac{1}{\gamma}\int_\Omega \left\|\tilde{A}\right\|_{L^1(\mathbb{R}^d)} \left\|\tilde{f}(x,\cdot,t)\right\|_{L^2(\mathbb{R}^d)} \|f(x,\cdot,t)\|_{L^2(\mathbb{R}^d)},\end{aligned}$$

where we used Young's convolution inequality in the last step and moreover defined $\tilde{f}(x,\tilde{w},t) = f(x,\frac{\tilde{w}}{\gamma},t)$. Notice that $\left\|\tilde{f}(x,\cdot,t)\right\|_{L^2(\mathbb{R}^d)} = \sqrt{\gamma}\,\|f(x,\cdot,t)\|_{L^2(\mathbb{R}^d)}$ and

$$\left\|\tilde{A}\right\|_{L^1(\mathbb{R}^d)} = A_0 \int_{\mathbb{R}^d} e^{-\beta|\tilde{v}|^2}\,d\tilde{v} = A_0 \left(\frac{\pi}{\beta}\right)^{d/2}.$$

Hence, we can conclude with Theorem 2.2 that there exist constants $C, c > 0$, such that

(4.38)

$$\begin{aligned}\|f(t)\|_{L^2(\Omega\times\mathbb{R}^d)} &\le c\left(\|f_0\|_{L^2(\Omega\times\mathbb{R}^d)}\right)\exp\left(C\int_0^t \|\nabla a(s)\|_{C_b^0} + \frac{1}{\tau} + \frac{1}{\sqrt{\gamma}}\left(\frac{\pi}{\beta}\right)^{d/2} ds\right)\\ &\le c\,\left(\|f_0\|_{L^2(\Omega\times\mathbb{R}^d)}\right)\exp\left(C\int_0^t \|\nabla a(s)\|_{C_b^0}\,ds\right) e^{\frac{T}{\tau\sqrt{\gamma}}} e^{T(\pi/\beta)^{d/2}}.\end{aligned}$$

In the next step, we investigate derivatives of f. For this purpose, we introduce the notation $\partial_j = \partial_{z_j}$, $j = 1,\ldots,2d$, where $z = (x_1,\ldots,x_d,v_1,\ldots,v_d)$. In particular, $\partial_j = \partial_{x_j}$ for $j = 1,\ldots,d$ and $\partial_j = \partial_{v_{j-d}}$ for $j = d+1,\ldots,2d$.

Taking the derivative ∂_j and taking L^2 inner product in $\Omega\times\mathbb{R}^d$ with $\partial_j f$ leads on the right-hand side of (4.32) with an application of Young's convolution inequality to

$$\begin{aligned}&\iint_{\Omega\times\mathbb{R}^d}\int_{\mathbb{R}^d} \partial_j\Big(A(w,v)f(x,w,t)\Big)\,dw\,\partial_j f(x,v,t)\,dx\,dv\\ &= \iint_{\Omega\times\mathbb{R}^d}\int_{\mathbb{R}^d} f(x,w,t)\;\partial_j A(w,v)\,dw\,\partial_j f(x,v,t)\,dx\,dv\\ &\qquad + \iint_{\Omega\times\mathbb{R}^d}\int_{\mathbb{R}^d} A(w,v)\;\partial_j f(x,w,t)\,dw\,\partial_j f(x,v,t)\,dx\,dv\\ &\le \frac{1}{\gamma}\int_\Omega \left\|\partial_j\tilde{A}\right\|_{L^1(\mathbb{R}^d)}\left\|\tilde{f}(x,\cdot,t)\right\|_{L^2(\mathbb{R}^d)}\|\partial_j f(x,\cdot,t)\|_{L^2(\mathbb{R}^d)}\,dx\\ &\qquad + \int_\Omega \frac{1}{\gamma}\left\|\tilde{A}\right\|_{L^1(\mathbb{R}^d)}\left\|\partial_j\tilde{f}(x,\cdot,t)\right\|_{L^2(\mathbb{R}^d)}\|\partial_j f(x,\cdot,t)\|_{L^2(\mathbb{R}^d)}\,dx\\ &\le \frac{1}{\sqrt{\gamma}}\left(\left\|\partial_j\tilde{A}\right\|_{L^1(\mathbb{R}^d)} + \left\|\tilde{A}\right\|_{L^1(\mathbb{R}^d)}\right)\left(\left\|\tilde{f}(\cdot,\cdot,t)\right\|_{L^2(\Omega\times\mathbb{R}^d)} + \left\|\partial_j\tilde{f}(\cdot,\cdot,t)\right\|_{L^2(\Omega\times\mathbb{R}^d)}\right)\times\\ &\qquad\qquad \times\left(\|\partial_j f(\cdot,\cdot,t)\|_{L^2(\Omega\times\mathbb{R}^d)} + \|f(\cdot,\cdot,t)\|_{L^2(\Omega\times\mathbb{R}^d)}\right)\\ &\le \frac{1}{\sqrt{\gamma}}\left(\left\|\tilde{A}\right\|_{W^{1,1}(\mathbb{R}^d)}\right)\|f(t)\|^2_{H^1(\Omega\times\mathbb{R}^d)}.\end{aligned}$$

where we summed over all $j = 1,\ldots,2d$ in the last inequality

Notice that $f(x,w,t)$ does not depend on v and therefore $\partial_j f(x,w,t) = 0$ for $j = d+1,\ldots,2d$. Further, it holds that $\left\|\partial_j\tilde{A}\right\|_{L^1(\mathbb{R}^d)}$ is a real number for all d and j. Hence, we conclude that there exists a constant $c > 0$ such that

$$\|f(t)\|_{H^1(\Omega\times\mathbb{R}^d)} \le c\,e^{T/\tau}\left(\|f_0\|_{H^1(\Omega\times\mathbb{R}^d)}\right)\exp\left(C\int_0^t \|\nabla a(s)\|_{C_b^1}\,ds\right)K_A,$$

where K_A is a constant depending on the KS kernel.

REMARK 4.7. *The structure $A(v,w) = \tilde{A}(v-w)$ is preserved for arbitrary derivatives with respect to v. Specifically, one can write $\partial_j^m A(v,w) = \partial_j^m \tilde{A}(v-w)$. Therefore, the strategy above for the H^1 case can be extended for H^m, $m \in \mathbb{N}$.*

4.2.3. A kinetic optimal control problem

In this section, we formulate our KS kinetic optimal control problem. We consider a finite time horizon $[0,T]$ and the phase space $\Omega \times \mathbb{R}^d$, where the spatial domain $\Omega \subset \mathbb{R}^d$ is bounded and convex with piecewise smooth boundary $\partial\Omega$. On this space, we consider the evolution of the density governed by (4.24). The Keilson-Storer collision kernel $A(v,w)$ is discussed in Section 4.1.

In order to ease notation, we denote with $C[f](x,v,t)$ the gain-loss collision term that appears on the right-hand side of (4.24). Further, we introduce the free-streaming operator L_u given by

$$L_u = v \cdot \nabla_x + u(x) \cdot \nabla_v.$$

Hence, our controlled kinetic model can be written as follows

$$\partial_t f(x,v,t) + L_u\, f(x,v,t) = C[f](x,v,t).$$

Next, we specify the boundary conditions. We have inflow boundary conditions for all $t \in (0,T]$, at $x \in \partial\Omega$, for all v for which it holds

$$v \in \mathbb{R}^d_< := \{v \in \mathbb{R}^d \,|\, v \cdot n(x) < 0\},$$

where $n(x)$ is the unit outward normal vector at $x \in \partial\Omega$. Then, choosing specular reflection, the boundary conditions at inflow are given by

$$f(x,v,t) = f(x, v - 2\,n\,(n \cdot v), t).$$

Moreover, we specify an initial density $f_0(x,v) \geq 0$ for all $(x,v) \in \Omega \times \mathbb{R}^d$ at time $t = 0$ such that

$$\lim_{|v|\to\infty} f_0(x,v) \to 0 \qquad \forall x \in \Omega. \tag{4.39}$$

Therefore, we have the following kinetic initial- and boundary-value problem

$$\begin{aligned} &\partial_t f(x,v,t) \,+\, L_u\, f(x,v,t) \,=\, C[f](x,v,t) && \text{in } \Omega \times \mathbb{R}^d \times (0,T] \\ &f(x,v,0) \,=\, f_0(x,v) && \text{on } \Omega \times \mathbb{R}^d \\ &f(x,v,t) = f(x, v - 2\,n\,(n \cdot v), t) && \text{on } \partial\Omega \times \mathbb{R}^d_< \times (0,T] \end{aligned} \tag{4.40}$$

Although a complete theoretical treatment of our kinetic control problem requires a separate work, for the present purpose we claim that for $T > 0$ fixed and $m \in \mathbb{N}$, assuming that $u \in H^1_0(\Omega) \cap C^{0,1}(\Omega)$ and choosing $f_0 \in H^m(\Omega \times \mathbb{R}^d)$ as specified above, then the problem (4.40) admits a unique weak solution $f \in C\Big([0,T]; H^m(\Omega \times \mathbb{R}^d)\Big)$. Moreover, supposing $f_0 \geq 0$, it holds that f is non-negative, and the property (4.39) is preserved along evolution since the control is assumed to be Lipschitz-continuous. Our claim for the case $m = 0$ follows from [**14**], Theorem 8. This claim for the case $m \in \mathbb{N}$ is underpinned by results of [**29**] for initial conditions in the space of tempered distributions and Remark 4.7.

Specifically, we claim that for the solution f it holds

$$\lim_{|v|\to\infty} f_0(x,v,t) \to 0 \qquad \text{for all} \quad (x,t) \in \Omega \times [0,T]. \tag{4.41}$$

We see that, choosing a fixed initial condition f_0, for a given control u the solution to (4.40) defines a control-to-state map $G : H_0^1(\Omega) \to C\Big([0,T]; H^m(\Omega \times \mathbb{R}^d)\Big)$ given by $u \mapsto f = G(u)$ that is well defined and continuous; see Chapter 2 for a proof in the case without collision and Remark 4.7 for arbitrary $m \in \mathbb{N}$ including collision following the strategies of Chapter 2.

Now, we discuss the formulation of our cost functional. As in [**12, 13**], we focus on a functional that appears in the framework of ensemble control problems [**12, 32, 33, 34**]. This choice is consistent with the significance of the density and kinetic equations in statistical mechanics. We have

$$J(f,u) = \int_0^T \int_{\Omega\times\mathbb{R}^d} \theta(z,t)\, f(z,t)\, dz\, dt + \int_{\Omega\times\mathbb{R}^d} \varphi(z) f(z,T)\, dz + \frac{\nu}{2}\, \|u\|_U^2 \,. \tag{4.42}$$

where $z = (x,v) \in \Omega \times \mathbb{R}^d$. Notice that the first and second terms in (4.42), can be understood as expected values by defining

$$\left\langle \int_0^T \theta(\cdot,t)\, dt \right\rangle = \int_0^T \int_{\Omega\times\mathbb{R}^d} \theta(z,t)\, f(z,t)\, dz\, dt$$

and

$$\langle \varphi(\cdot) \rangle = \int_{\Omega\times\mathbb{R}^d} \varphi(z)\, f(z,T)\, dz.$$

We refer to the first term as the tracking term (also called integrated cost) and to the second term as the final observation (also called terminal cost). In addition, we have the cost of the control with weight $\nu > 0$. With U we denote the Hilbert space where the control is sought, and $\|\cdot\|_U$ is the corresponding norm.

In order to illustrate the role and name of θ in the cost functional, suppose that $\zeta_D(t)$ represents a desired mean position-velocity profile for the ensemble of our particles. Then, the choice $\theta(z,t) = |z - \zeta_D(t)|^2$ results in a tracking term that achieves its minimum if all particles exactly follow the desired trajectory ζ_D such that the density is concentrated on it. Similarly, the final observation term can be defined as $\varphi(z) = |z - \zeta_T|^2$, which corresponds to the requirement that at final time the particles get close to ζ_T. In general, we make the following

Assumption 4.2. *We suppose that θ and φ are integrable smooth functions, bounded from below and locally convex in a neighbourhood of ζ_D and ζ_T, so that the minus gradient of $\theta(\cdot, t)$, resp. $\varphi(\cdot)$, is always pointing to the unique global minimum of the respective functions for all $t \in [0, T]$.*

Concerning the third term in (4.42), we choose $U = H_0^1(\Omega)$ as our control space. Therefore,

$$\|u\|_U^2 = \int_\Omega |u(x)|^2 \, dx + \int_\Omega |\nabla u(x)|^2 \, dx.$$

This choice is motivated by the requirement of a minimal degree of regularity of our control force such that it could be implemented in laboratory. As a modelling choice, we require that the force is zero at the boundary of the space domain (other choices are possible). Notice that the value of the weight $\nu > 0$ establishes the relative importance of achieving the given target with respect to the cost of the corresponding control.

Now, we can formulate our kinetic optimal control problem

$$\begin{aligned} \min \ J(f,u) := \int_0^T \int_{\Omega\times\mathbb{R}^d} \theta(x,v,t)\, f(x,v,t)\, dx\, dv\, dt \\ + \int_{\Omega\times\mathbb{R}^d} \varphi(x,v) f(x,v,T)\, dx\, dv \, + \, \frac{\nu}{2}\, \|u\|_U^2 \end{aligned}$$

$$\text{s.t.} \begin{cases} \partial_t f(x,v,t) + L_u\, f(x,v,t) = C[f](x,v,t), & \text{in } \Omega \times \mathbb{R}^d \times (0,T] \\ f(x,v,0) = f_0(x,v) & \text{on } \Omega \times \mathbb{R}^d \\ f(x,v,t) = f(x, v - 2n(n\cdot v), t) & \text{on } \partial\Omega \times \mathbb{R}^d_< \times (0,T] \end{cases} \tag{4.43}$$

$$u \in U.$$

By means of the control-to-state map, this constrained optimization problem can be reformulated as the following unconstrained minimization problem

$$\min_{u\in U} J_r(u), \tag{4.44}$$

where $J_r(u) := J(G(u), u)$ defines the so-called reduced cost functional.

In this setting, we can state existence of a minimizer (i.e. an optimal control) to (4.44) as follows.

Theorem 4.2. *Let θ and φ fulfil Assumption 4.2 and let $f_0 \in H^m(\Omega \times \mathbb{R}^d)$, $m \in \mathbb{N}$. Then the linear kinetic optimal control problem* (4.43) *has at least one solution $u^* \in U = H_0^1(\Omega)$ with corresponding optimal state $f \in C\Big([0,T]; H^m(\Omega \times \mathbb{R}^d)\Big)$.*

Proof. The functional J given in (4.42) is well-defined for $(f,u) \in C\Big([0,T]; H^m\Big) \times H_0^1(\Omega)$. It is bounded from below and coercive in u; consequently, J_r is bounded. Further, J_r is weakly lower semi-continuous since G is continuous, the last term is a norm, so it is weakly lower semi-continuous, and the first two terms are linear in f, and then they are weakly continuous. Thus, we have that, if $\left(u^\kappa\right)_\kappa \subset U$ is a sequence which converges weakly to a $u \in U$, we have

$$\liminf_{\kappa\to+\infty} J_r(u^\kappa) \;=\; \liminf_{\kappa\to+\infty} J\Big(G(u^\kappa), u^\kappa\Big) \;\geq\; J\Big(G(u), u\Big) \;=\; J_r(u)\,.$$

At this point, proving the existence of a minimizer for J_r is standard: Let us take a minimizing sequence $\left(u^\kappa\right)_\kappa \subset U$. This sequence is bounded by the coerciveness of the cost functional in u. Since U is a Hilbert space, we can extract a weakly convergent subsequence, which we do not relabel for simplicity; let us call $u^* \in U$ its limit-point. Then, by the weak lower semi-continuity of J_r, we can conclude that u^* is a minimizer. □

We remark that in our kinetic model (4.24) the control force u multiplies $\nabla_v f$, and this bilinear structure makes our optimization problem non-convex and nonlinear. For this reason, in general, it is not possible to establish uniqueness of the minimizer u^*.

4.2.4. Kinetic optimality system

In this section, we discuss the optimality system characterizing a solution to (4.43). Since (4.43) and (4.44) are equivalent, and assuming Fréchet differentiability of G and J, in the absence of control constraints, the first-order necessary optimality condition for a solution to (4.44) is given by

$$\nabla_u J_r(u) = 0.$$

Notice that the computation of $\nabla_u J_r(u)$, that is, of $\nabla_u J(G(u), u)$, involves the Fréchet differentiability of G and J, which we assume in this chapter; however, see Chapter 2, where we prove Fréchet differentiability for a similar problem, and Remark 4.2.

A convenient way to derive $\nabla_u J_r(u)$ is to introduce the Lagrange function corresponding to (4.43) as follows

$$\begin{aligned}\mathcal{L}(f, u, q, q_0, q_\sigma) := J(f, u) + \int_0^T \int_{\Omega\times\mathbb{R}^d} \Big(\partial_t f(z,t) + L_u f(z,t) - C[f](z,t)\Big)\, q(z,t)\, dz\, dt \\ + \int_{\Omega\times\mathbb{R}^d} \Big(f(z,0) - f_0(z)\Big)\, q_0(z)\, dz \\ + \int_{L^2(\partial\Omega\times\mathbb{R}^d_<\times[0,T])} (f(x,v,t) - f(x, v - 2n(n\cdot v), t))\, q_\sigma(x,v,t)\, d\sigma\, dv\, dt,\end{aligned}$$

where $z = (x, v)$, q, q_0 and q_σ represent Lagrange multipliers and $d\sigma$ is a surface element. In this framework, the so-called optimality system is obtained by requiring that the Fréchet derivatives of $\mathcal{L}$ with respect to each of its arguments are zero.

The resulting optimality system consists of three parts: 1) the kinetic model; 2) the adjoint kinetic problem; 3) the optimality condition that corresponds to $\nabla_u J_r(u) = 0$, and is expressed in terms of the solutions to the two problems in 1) and 2).

For the adjoint model, we obtain

$$\begin{aligned} -\partial_t q(x,v,t) + L_u^* \, q(x,v,t) = \int_{\mathbb{R}^d} A(v,w)\, q(x,w,t)\, dw \\ -q(x,v,t) \int_{\mathbb{R}^d} A(v,w)\, dw - \theta(x,v,t) \qquad & \text{in } \Omega \times \mathbb{R}^d \times (0,T], \\ q(x,v,T) = -\varphi(x,v) \qquad & \text{in } \Omega \times \mathbb{R}^d, \\ q(x,v,t) = q(x, v - 2n(n\cdot v), t) \qquad & \text{in } \partial\Omega \times \mathbb{R}^d_{>} \times [0,T]\ , \end{aligned} \tag{4.45}$$

where $\mathbb{R}^d_{>} := \{v \in \mathbb{R}^d \,|\, v \cdot n(x) > 0\}$. Specifically, the adjoint free-streaming operator L_u^* is derived through integration by parts and is given by

$$L_u^* := -v \cdot \nabla_x - u(x) \cdot \nabla_v = -L_u.$$

Now, as in the last section, we are confronted with the problem that (4.45) has not the structure of a kinetic equation, that is, the right-hand side of the adjoint equation cannot be interpreted as a gain-loss term. We recover the required structure as done in Section 4.1.2, in which C_0^* has been defined. Analogously, we define the uncontrolled adjoint collision term

$$C^*[q](x,v,t) = \int_{\mathbb{R}^d} A^*(w,v)\, q(x,w,t)\, dw - q(x,v,t) \int_{\mathbb{R}^d} A^*(v,w)\, dw,$$

where $A^*(w,v) = \frac{1}{\gamma} A(v,w)$, which gives an 'adjoint' mean free time $\tau_q = \gamma\, \tau$.

Summarizing, we can write the adjoint equation as follows

$$-\partial_t q(x,v,t) + L_u^* q(x,v,t) = C^*[q](x,v,t) + C_0^*\, q(x,v,t) - \theta(x,v,t).$$

This is a linear kinetic equation with, in addition, a linear reaction term and a source term. For this equation, a terminal condition is given and therefore we consider the adjoint variable evolving backwards in time.

Next, to conclude the formulation of the optimality system, we discuss the optimality condition. From the Lagrange function, we obtain

$$-\nu\, \Delta u(x) + \nu\, u(x) + \int_0^T \int_{\mathbb{R}^d} q(x,v,t)\ (e \cdot \nabla_v f(x,v,t))\ dv\, dt = 0, \tag{4.46}$$

where we have defined the vector $e = (1, \ldots, 1)^T \in \mathbb{R}^d$ and $\Delta u = \sum_{i=1}^n \partial^2_{x_i x_i} u$. This equation defines an elliptic problem for u where homogeneous Dirichlet boundary conditions are required. Notice that the left-hand side of (4.46) represents the L^2 gradient $\nabla_u J_r(u)$.

Summarizing, our kinetic optimality system is given by

$$\begin{aligned}
&\partial_t f(x,v,t) + L_u f(x,v,t) = C[f](x,v,t)\\
&f(x,v,0) = f_0(x,v)\\
&f|_{\partial\Omega\times\mathbb{R}^d_<} = f(x, v-2n(n\cdot v),t)\\
&\\
&-\partial_t q(x,v,t) + L_u^* q(x,v,t) = C^*[q](x,v,t) + C_0^*\, q(x,v,t) - \theta(x,v,t)\\
&q(x,v,T) = -\varphi(x,v)\\
&q|_{\partial\Omega\times\mathbb{R}^d_>} = q(x,v-2n(n\cdot v),t)\\
&\\
&-\Delta u(x) + u(x) = -\frac{1}{\nu}\int_0^T\int_{\mathbb{R}^d} q(x,v,t)\,(e\cdot\nabla_v f(x,v,t))\,dv\,dt\\
&u|_{\partial\Omega} = 0.
\end{aligned}$$

Our aim is to solve this system using a Monte Carlo methodology and a gradient-based optimization scheme. For this reason, we would like to interpret q as a density function and, correspondingly, we refer to adjoint particles. We remark that the main purpose of the adjoint variable q is to allow the computation of the reduced gradient. Therefore, also a suitable approximation of q that results in an effective gradient is appropriate for our goal.

For convenience of implementation, we consider the following transformation for the velocity field $\bar{v} = -v$, which leads to

$$\begin{aligned}
&-\partial_t q(x,\bar{v},t) - L_u^* q(x,\bar{v},t) = C_u^*[q](x,\bar{v},t) + C_0^* q(x,\bar{v},t) - \theta(x,\bar{v},t),\\
&q(x,\bar{v},T) = -\varphi(x,\bar{v})\\
&q|_{\partial\Omega\times\mathbb{R}^d_<} = q(x,\bar{v}-2n(n\cdot\bar{v}),t).
\end{aligned}\tag{4.47}$$

Notice that the boundary condition is now formulated on $\partial\Omega\times\mathbb{R}^d_<\times[0,T]$. Also for this adjoint problem, we claim existence and uniqueness of a non-negative solution, assuming the following choice of the functions θ and φ. We have

$$\theta(z,t) = -\frac{C_\theta}{\sqrt{(2\pi)^{2d}\det(\Sigma_\theta)}}\exp\left(-\frac{1}{2}(z-\zeta_D(t))^T\Sigma_\theta^{-1}(z-\zeta_D(t))\right),\qquad C_\theta > 0,$$

where C_θ represents a weight of the tracking part of the cost functional, and $\Sigma_\theta \in \mathbb{R}^{2d\times 2d}$ has the significance of a co-variance matrix that we assume to be a diagonal matrix.

This choice for θ is motivated by the fact that we can implement the corresponding source term in (4.47) in a MC framework by adding adjoint particles to the distribution q in every time-step based on the Gaussian distribution given by $-\theta$.

Thus, at each time-step, we add a certain number of particles N_{frac} obeying the $2d$-dimensional multi-variate Gaussian distribution $\mathcal{N}_{2d}(\zeta_D(t), \Sigma)$ with mean $\zeta_D(t)$ and co-variance matrix Σ.

Similarly, we choose

$$\varphi(z) = -\frac{C_\varphi}{\sqrt{(2\pi)^{2d}\det(\Sigma_\varphi)}} \exp\left(-\frac{1}{2}(z-\zeta_T)^T \Sigma_\varphi^{-1}(z-\zeta_T)\right), \qquad C_\varphi > 0.$$

Notice that the choice of θ and φ given above satisfy the requirements of Assumption 4.2.

We conclude this section with some theoretical consideration for the case $d = 1$ and $\Omega \subset \mathbb{R}$. Assume that for the initial guess of the control it holds $u^0 \in C^1(\Omega)$ with homogeneous Dirichlet boundary condition, assume that $f_0 \in H^1(\Omega \times \mathbb{R})$ and let θ and φ fulfil Assumption 4.2. Then it follows that any gradient based numerical optimization scheme provides a sequence of controls $(u^\kappa)_\kappa$ such that $u^\kappa \in C^1(\Omega)$ for all $\kappa \in \mathbb{N}$. Indeed, this is proven by induction over $\kappa \in \mathbb{N}_0$.

By assumption it holds that $u^0 \in C^1(\Omega)$. Assume now $u^\kappa \in C^1(\Omega)$ for an arbitrary but fixed $\kappa \in \mathbb{N}$. Then there exists a unique solution with H^1-regularity for the forward and backward models with the control $u = u^\kappa$. Since $u^{\kappa+1}$ is obtained by solving the second-order elliptic partial differential equation (4.46), we can apply certain regularity theorems. Specifically, since the integral in equation (4.46) is in $L^2(\Omega)$, we can conclude that for the solution holds $u^{\kappa+1} \in H^2(\Omega)$, see [**66**]. In one dimension, we further have the Sobolev embedding $H^2 \hookrightarrow C^1$; see [**1, 66**].

4.2.5. A Monte Carlo scheme in phase space and numerical optimization

In this section, we illustrate a MC scheme for solving our kinetic control problem. We show how to adjust the methods of Section 4.1.3. We focus on the case $d = 1$ and choose $\Omega = (0, L)$, $L > 0$; however, our methodology applies analogously in higher dimension. Keep in mind, that for $d \geq 2$, one has more degrees of freedom for the reflecting boundary and has to assume higher regularity of the initial condition and of the initial guess for the control.

Next, we discuss our implementation of the free-streaming operator L_u. The free streaming time is given by δt, calculated using (4.18). Within this time lapse, the microscopic equations of motion have to be integrated, and for this purpose we apply the (velocity) Verlet algorithm; see, e.g., [**82, 115, 130**]. An important property of this method is that it is symplectic which means that it is volume-preserving in phase-space.

While updating the position, one has to take the boundedness of the physical domain and the boundary condition into account. For this purpose, the distance between the updated position and the boundary of Ω is important. Since the particles are considered in the mean inside Ω, this distance is connected to the physical property that the particle can cross the domain several times. To implement the boundary condition, we present Algorithm 4.8.

Algorithm 4.8 Kinetic model, boundary condition

```
Require: Updated position $\tilde{x}$ according to Verlet method
if $\tilde{x} \in \Omega = [0, L]$ then
    return $\tilde{x}$
else if $\tilde{x} < 0$ then
    $\omega = \lfloor \tilde{x}/L \rfloor \mod 2$
    set velocity $v = (-1)^{\omega-1} v$
    return $x = \omega L + (-1)^{\omega}(-\tilde{x} \mod L)$
else if $\tilde{x} > L$ then
    $\omega = \lfloor \tilde{x}/L \rfloor \mod 2$
    set velocity $v = (-1)^{\omega-1} v$
    return $x = (1-\omega)L + (\omega - 1)(\tilde{x} \mod L)$
end if
```

At this point, we can illustrate our MC kinetic model solver with the following algorithm, where the initial condition f_0 is used to initialize the list of particles (pointers) in the sense that it provides the density of the initial distribution of the velocities of the particles. In the initialization, we choose a number of particles N_f. Further, we consider a partition of the time interval $[0, T]$ in N_t subintervals of size $\Delta t = T/N_t$ such that $\Delta t \gg \delta t$. With this setting, we have $t^k = k\Delta t$, for the time of the k-th time-step, $k = 0, \ldots, N_t$.

In our implementation, we define F as the list of labelled pointers to structures that resemble particles. We denote with $F^k[p]$ the pointer to the p-th particle at the k-th time-step. We have $p = 1, \ldots, N_f$ and $k = 0, \ldots, N_t$. Further, let $F^k[p].v$ be the velocity of the p-th particle at the k-th time-step, and let $F^k[p].x$ be the position of the p-th particle at the k-th time-step.

Moreover, let $F^k[p].t'$ be the time that is elapsed for the p-th particle starting from t^k. This quantity is used to determine if the particle will undergo another collision in the current time-step, assuming that $0 \leq F^k[p].t' < \Delta t$. Analogously, we denote with Q the list of labelled pointers to structures representing adjoint particles.

To initialize F^0 using the distribution f_0, we apply Algorithm 4.9 given below, which is analogue to 4.1 except from the dependence on the position. A similar algorithm applies to initialize Q^{N_t} with the distribution $-\varphi$.

Algorithm 4.9 Generation of initial condition

```
Require: $f_0(x, v)$
for $p = 1$ to $N_f$ do
    Compute $(F^0[p].v, F^0[p].x) \sim f_0(x, v)$
    Set $F^0[p].t' = 0$
end for
```

Our Monte Carlo kinetic model solver is implemented as follows.

Algorithm 4.10 Monte Carlo kinetic model solver

```
Require: $f_0(x, v)$, $u(x)$
Initialise $N_f$ particles using Algorithm 4.9 and $f_0(x, v)$, set $\delta t_2 = 0$
for $k = 0$ to $N_t - 1$ do
    for $p = 1$ to $N_f$ do
        while $F^k[p].t' < \Delta t$ do
            Compute $\delta t_1$ according to (4.18)
            Determine $F^k[p].v \sim \mathcal{N}\left(\gamma v, \frac{1}{2\beta}\right)$
            update $F^k[p].x$ and $F^k[p].v$ according to the Verlet-Algorithm:
            $F^k[p].x \; = \; F^k[p].x + F^k[p].v \, \delta t_1 \; + \; u(F^k[p].x) \frac{\delta t_1 + \delta t_2}{2} \delta t_1,$
            $F^k[p].v \; = \; F^k[p].v + u(F^k[p].x) \, \delta t_1$
            and taking the boundary condition into account using Algorithm 4.8
            $F^k[p].t' = F^k[p].t' + \delta t_1$
            $\delta t_2 = \delta t_1$
        end while
        if $F^k[p].t' > \Delta t$ then
            $F^{k+1}[p].t' = F^k[p].t' \bmod \Delta t$
        end if
    end for
end for
```

Analogous to the treatment of the source and linear reaction term in Section 4.1, we present the following two algorithms. They are analogous to 4.3 and 4.4, respectively.

Algorithm 4.11 Implementation of the source term θ at time t^k

```
Require: $\zeta_D(t^k)$, $\Sigma_\theta$, $N_{frac}$
Generate $N_{frac}$ new particles with velocity and position components having the
    normal distribution with mean $\zeta_D(t^k)$ and variance $\Sigma_\theta$: $(x, v) \sim \mathcal{N}\left(\zeta_D(t^k), \Sigma_\theta\right)$
Add these particles to the existing ones in $Q^k$
```

Algorithm 4.12 Implementation of the linear reaction term at time t^k

```
Require: $Q^k$, $N_q^k$
Set $N \in \mathbb{N}$, $\varepsilon \in [0,1)$ such that $\Delta t\, C_0^* = N + \varepsilon$
for $p = 1$ to $N_q^k$ do
    Generate $N$ particles with the velocity and position $(Q^k[p].x, Q^k[p].v)$
    Generate uniform random number $r \in [0,1]$
    if $r > 1 - \varepsilon$ then
        Generate a particle with the velocity and position $(Q^k[p].x, Q^k[p].v)$
    end if
end for
Add generated particles to the existing ones in $Q^k$
```

Notice that, since in the implementation of the adjoint kinetic model we vary the number of adjoint particles depending on the linear reaction term and the source term, we index this number with k and write N_q^k. In Algorithm 4.11, we choose $N_{frac} \ll N_f$.

With these two procedures, we can implement the time evolution of the adjoint variable starting from the terminal condition given by $-\varphi(x,v)$. This function is used to initialize the list of adjoint particles (pointers) in the sense that it provides the density of the initial distribution of the velocities of these particles.

Algorithm 4.13 Monte Carlo adjoint kinetic model solver

```
Require: $\theta(x,v,t)$, $\varphi(x,v)$, $u(x)$.
Initialize $Q^{N_t}$ with $N_q^{N_t} = N_{frac}$ particles using Algorithm 4.9 and $-\varphi$, set $\delta t_2 = 0$

for $k = N_t$ to 1 do
    Use Algorithm 4.11 to implement the source term
    Use Algorithm 4.12 to implement the linear reaction term
    for $p = 1$ to $N_q^k$ do
        while $Q^k[p].t' < \Delta t$ do
            Generate $\delta t_1$ according to (4.18) using $\tau_q$ instead of $\tau$
            Determine $v \sim \mathcal{N}\left(\frac{v}{\gamma}, \frac{1}{2\beta\gamma^2}\right)$
            update $Q^k[p].x$ and $Q^k[p].v$ according the adjoint Verlet-Algorithm:
            $Q^k[p].x \;=\; Q^k[p].x + Q^k[p].v\,\delta t_1 \;-\; u(Q^k[p].x)\frac{\delta t_1+\delta t_2}{2}\delta t_1$,
            $Q^k[p].v \;=\; Q^k[p].v - u(Q^k[p].x)\,\delta t_1$, and taking the adjoint boundary condition into account using Algorithm 4.8
            $Q^k[p].t' = Q^k[p].t' + \delta t$
            $\delta t_2 = \delta t_1$
        end while
        if $Q^k[p].t' > \Delta t$ then
            $Q^{k-1}[p].t' = Q^k[p].t' \bmod \Delta t$
        end if
    end for
end for
```

Now, analogue to Section 4.1.3, we define

$$\Upsilon_{\Delta v} := \{ \quad v_j \in \Upsilon, \quad j = 1, \dots, N_v \quad \}, \qquad v_j := \left(j - \frac{1}{2}\right) \Delta v - v_{\max}$$

and analogously the spatial domain

$$\Omega_{\Delta x} := \{ \quad x_i \in \Omega, \quad i = 1, \dots, N_x \quad \}, \qquad x_i := \left(i - \frac{1}{2}\right) \Delta x.$$

Hence, we have the discretized phase space $\Omega_{\Delta x} \times \Upsilon_{\Delta v}$. On the other hand, we recall that on the time interval $[0, T]$, we have the time-steps $t^k := k\Delta t$, $k = 0, \dots, N_t$, and define

$$\Gamma_{\Delta t} := \left\{ \quad t^k \in [0, T], \quad k = 0, \dots, N_t \quad \right\}.$$

Now, we denoted with f_{ij}^k the occupation number of the cell centred in (x_i, v_j) in the phase space. To construct this function, we count the particles at time-step k that have position and velocity in the cell centred at (x_i, v_j). Thus, we define analogue to (4.20)

$$f_{ij}^k := \sum_{p=1}^{N_f} \mathbb{1}_{ij} \left(F^k[p].x, F^k[p].v \right). \tag{4.48}$$

It results that, if a particle with position x and velocity v within a cell centred at $(x_i, v_j) \in \Omega \times \mathbb{R}$ is subject to collision and acquires a new velocity v' within a cell centred at $(x_i, v_k) \in \Omega \times \mathbb{R}$, then the value of f_{ij} is reduced by 1 and, on the other hand, the value of f_{ik} is increased by 1. Notice that choosing $v_{\max}$ large enough, the probability that the velocity of a particle exceeds the boundary of Υ after collision is very low but possibly not zero. If this rare event happens, we generate again a new velocity for the particle using the same pre-collision velocity as before.

Therefore, we only need to consider the adjoint particles discrete distribution analogue to 4.21

$$q_{ij}^k = \sum_{p=1}^{N_q^k} \mathbb{1}_{ij} \left(Q^k[p].x, Q^k[p].v \right). \tag{4.49}$$

Since we have a finite number of particles, the data that we obtain from the MC procedure method appears subject to noise. For this reason and to facilitate the construction of the gradient, we introduce a denoising procedure that can be put in the framework of Tikhonov (resp. Sobolev) regularization techniques, see also [**93**]. Thus, denoising can be interpreted as solving the minimization problem

$$\min_{f \in H^1(\Omega \times \Upsilon \times [0,T])} \frac{c_s}{2} \int_\Upsilon |\nabla f|^2 \, dv + \frac{1}{2} \int_\Upsilon \left(f - \tilde{f} \right)^2 \, dv, \tag{4.50}$$

where $\tilde{f}$ is the original noisy data obtained by assembling and $c_s > 0$ is a regularization parameter that is usually small compared to the maximum value of $\tilde{f}$.

The Euler-Lagrange equation corresponding to the optimization problem (4.50) is given by

$$-c_s \Delta f + (f - \tilde{f}) = 0, \tag{4.51}$$

where $\Delta = \partial^2_{vv}$, and we impose homogeneous Neumann boundary conditions. In this way, our denoising technique conserves the total mass. To show this fact, consider the integration of (4.51). We have

$$-c_s \int_\Upsilon \Delta f \, dv \, + \, \int_\Upsilon f \, dv \, = \, \int_\Upsilon \tilde{f} \, dv.$$

Further, by the divergence theorem and the given boundary conditions, we get that the resulting surface integral over $\partial\Upsilon$ vanishes. Thus, we obtain

$$\int_\Upsilon f \, dv \, = \, \int_\Upsilon \tilde{f} \, dv.$$

The linear elliptic problem given by (4.51) and homogeneous Neumann boundary conditions can be solved numerically. For this purpose, we put it in the form

$$\left(I - c_s D_2^\pm\right) f = \tilde{f},$$

where $D_2^\pm$ is the standard discretized second-order derivative in velocity space. The solution of this algebraic problem can be computed by standard methods of numerical linear algebra.

Next, we focus on the gradient equation (4.46) and assemble the L^2 optimization gradient corresponding to the left-hand side of (4.46) in the vector $g \in \mathbb{R}^{(N_x-2)}$. We use a rectangular quadrature rule to approximate the integrals in (4.46) and obtain

$$G_i := (\Delta v)(\Delta t) \left[\sum_{k=1}^{N_t} \sum_{j=0}^{N_v-1} \frac{f^k_{i,j+1} - f^k_{ij}}{\Delta v} q^k_{ij} \right] \qquad i = 2, \ldots, N_x - 1,$$

Further, to formulate the discrete version of (4.46) we use finite differences. Therefore, our L^2 gradient is given by

$$g_i := \nu u_i - \frac{\nu}{\Delta x^2} \left(u_{i+1} - 2u_i + u_{i-1}\right) + G_i \qquad i = 2, \ldots, N_x - 1.$$

Since our control field is required in $H^1_0(\Omega)$, we need to develop the H^1 representation of our gradient. For this purpose, we present the following reasoning that illustrates how to arrive at this representation.

Consider a Taylor expansion of the reduced cost functional $J_r(u)$ in a Hilbert space X as follows

$$J_r(u + \epsilon \, \delta u) = J_r(u) + \epsilon \, (\nabla J_r(u), \delta u)_X + \frac{\epsilon^2}{2} \left([\nabla^2 J_r(u)]\delta u, \delta u\right)_X + O(\epsilon^3) \tag{4.52}$$

The actual gradient depends on the choice of which inner product space we use. If we choose the space $X = L^2(\Omega)$, we have the inner product $(u, v)_X = \int_\Omega u(x) \cdot v(x) \, dx$

and the gradient is given by

$$\nabla J_r(u)|_{L^2} = \nu\, u(x) - \nu\, u''(x) - \int_0^T \int_{\mathbb{R}} q(x,v,t)\, \nabla_v f(x,v,t)\, dv\, dt. \tag{4.53}$$

In the case of $X = H^1(\Omega)$, we can determine the H^1 gradient based on the fact that the Taylor series must be identical term-by-term regardless of the choice of X. Therefore, we have

$$(\nabla J_r(u)|_{H^1}, \delta u)_{H^1} = (\nabla J_r(u)|_{L^2}, \delta u)_{L^2} \,.$$

Using the definition of the H^1 inner product $(u,v)_{H^1} = (u,v)_{L^2} + (u',v')_{L^2}$, we obtain the relation

$$\int_\Omega \Big(\nabla J_r(u)|_{H^1}(x)\, \delta u(x) + \frac{\mathrm{d}}{\mathrm{d}x} \nabla J_r(u)|_{H^1}(x)\, \delta u'(x) \Big)\, dt = \int_\Omega \nabla J_r(u)|_{L^2}(x)\, \delta u(x)\, dx,$$

which must hold for all test functions δu. Integrating by parts the second term in the integral on the left-hand side, with the assumption that the control is zero at $x = 0$ and $x = L$, we obtain the following equation for the H^1 gradient.

$$-\frac{\mathrm{d}^2}{\mathrm{d}x^2}[\nabla J_r(u)|_{H^1}(x)] + [\nabla J_r(u)|_{H^1}(x)] = \nabla J_r(u)|_{L^2}(x), \tag{4.54}$$

with the conditions $J_r(u)|_{H^1}(0) = 0$ and $J_r(u)|_{H^1}(L) = 0$. We approximate this problem by standard finite difference approximation, which results in a block-tridiagonal system. The solution of this system is efficiently obtained by the Thomas method.

With this preparation, we can formulate the algorithm that provides the $J_r(u)|_{H^1}(x)$ gradient that is required in our optimization scheme.

Algorithm 4.14 Calculate the gradient $\nabla J_r(u)|_{H^1}(x)$

```
Require: control $u(x)$, $f_0(x,v)$, $\varphi(x,v)$, $\theta(x,v,t)$
1: Solve the kinetic problem using Algorithm 4.10 with inputs $f_0(x,v)$, $u(x)$
2: Solve adjoint kinetic problem using Algorithm 4.13 with inputs $\varphi(x,v)$, $\theta(x,v,t)$,
   $u(x)$
3: Determine the distributions $f$ and $q$ according to (4.48) and (4.49)
4: Smoothing of the distributions $f$ and $q$
5: Assemble $\nabla J_r(u)|_{L^2}(x)$ according to (4.53)
6: Compute $\nabla J_r(u)|_{H^1}(x)$ solving (4.54)
```

We remark that, with this algorithm, we can implement many different gradient-based optimization schemes [**28**]. In our case, we choose the non-linear conjugate gradient (NCG) method. This is an iterative method that constructs a minimizing sequence of control functions (u^n) as illustrated by the following algorithm, analogue to 4.7.

Algorithm 4.15 NCG scheme

```
Require: $u^0(x)$, $f_0(x,v)$, $\varphi(x,v)$, $\theta(x,v,t)$
$n = 0$, $E > tol$
Compute $h^0 = -\nabla J_r(u^0)|_{H^1}$ using Algorithm 4.14
while $E > tol$ and $n < n_{\max}$ do
    Use a line-search scheme to determine the step-size $\alpha_n$ along $h^n$
    Update control: $u^{n+1} = u^n + \alpha_n\, h^n$
    Compute $d^{n+1} = \nabla J_r(u^{n+1})|_{H^1}$ using Algorithm 4.14
    Compute $\beta_n$ using the Fletcher-Reeves formula
    Set $h^{n+1} = -d^{n+1} + \beta_n\, h^n$
    $E = \|u^{n+1} - u^n\|$
    Set $n = n + 1$
end while
return $(u^n, f^n)$
```

In this algorithm, the tolerance $tol > 0$ and the maximum number of iterations $n_{\max} \in \mathbb{N}$ are used as termination criteria. We use backtracking line-search with Armijo condition. The factor β_n is based on the Fletcher and Reeves formula; see [**28**] for more details and references.

4.2.6. Numerical experiments

In this section, we perform numerical experiments in an one-dimensional position domain and one-dimensional velocity domain.

We implement a MOCOKI (Monte Carlo approach for optimal control in the force of a linear kinetic model) code. The MOCOKI code is a standalone `C++` package that realizes the Algorithm 4.15. The numerical and optimization parameters can be specified in the file *globalparameters.h*. The most important constants are given in Table 4.2. In *globalparameters.h* the purpose of each parameter is explained. The file *optimization/optimization_algorithms.cpp* is the core of the optimization process and mirrors Algorithm 4.15 of this section. In this file, the initial condition for the kinetic model and the optimization functions can be specified. After coding all the required parameters and functions, the code should be compiled using `cmake` and can be run using the command `./MOCOKI` in the console. During execution, the code generates several csv and txt files containing information on all particles, the value of the objective, the norm of the gradient, and the control at every optimization iteration.

After completion, it is possible to execute the python file `solution_kinetic_model.py` in the folder *post_processing*, which will result in displaying six plots. These figures show the evolution of the mean position and velocity in phase space, these values plus and minus the corresponding variance, the evolution of position and velocity over the time interval, the control together with the force used to generate the desired trajectory as a single particle dynamical system, and the development of the relative value of the functional during the calculation.

For a first experiment, we use the the evolution equations of the moments of linear kinetic models with the Keilson-Storer collision kernel derived in Section 4.2.1 to validate our MOCOKI code. Afterwards, we present a verification strategy for our MOCOKI implementation for a simple test-case. Specifically, we want to gather all the particles in the centre of the position domain. The last part of this section is to validate our implementation with standard, more evolved test-case of a harmonic oscillator. Also for this test-case the optimization procedure finds a control that is close to the well-known theoretical force of an ideal harmonic oscillator.

We now validate our code using the equations that were found for the evolution of the moments for the linear kinetic model (4.24); specifically, the equations (4.27), (4.30), (4.31). For the discrete versions of the moments we use a second-order trapezoidal quadrature rule for numerical integration. For discretizing the phase-space-time cylinder, we use the following parameters

$$\Delta t = \frac{T}{N_t} = 1.25 \cdot 10^{-3}, \qquad \Delta x = \frac{L}{N_x} = \frac{10}{200} = 0.05,$$
$$\Delta v = \frac{2v_{\max}}{N_v} = \frac{8}{400} = 0.02.$$

To calculate the moments, we compute the solution of (4.24) using Algorithm 4.10. To start it, we use an initial distribution $f_0(x,v)$ that is uniform in position and Gaussian in velocity. Specifically,

$$f_0(x,v) = \frac{1}{L}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{v^2}{2\sigma^2}\right).$$

Where $\sigma = 1/\sqrt{2\beta}$ where β is the parameter within the Keilson-Storer collision kernel. As control, we use

$$u(x) = 25 \sin\left(\frac{2\pi}{L}x\right).$$

After the execution of Algorithm 4.10 with this inputs, we calculate the moments and the corresponding discrete error functions as follows. Recall, that the moments are functions of space and time. We define the discretized versions of the number density as $(n_f)_i^k := n_f(x_i, t^k)$ and analogue for the higher moments. For the zeroth moment, we have

$$e_{n_f}^{i,k} := \frac{1}{\Delta t}\left((n_f)_i^{k+1} - (n_f)_i^k\right) + \frac{1}{2\Delta x}\left(\bar{m}_{i+1}^k - \bar{m}_{i-1}^k\right),$$

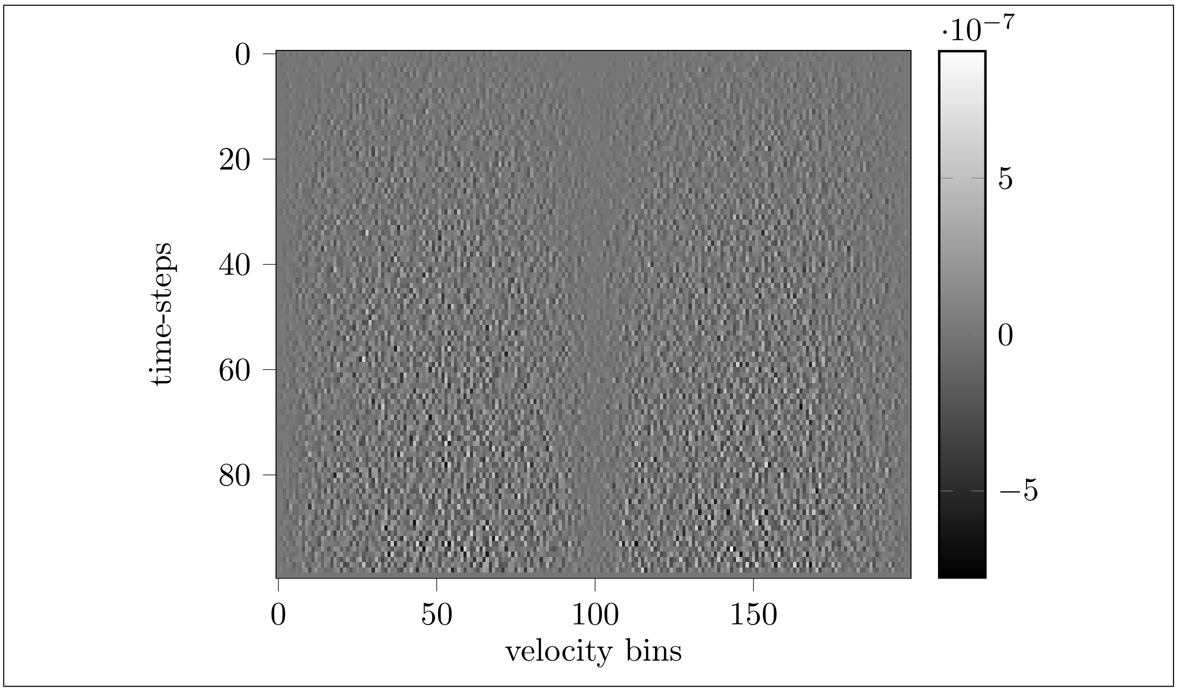


Figure 4.6. Error $e^{i,k}_{n_f}$ of zeroth moment.

for the first moment

$$e^{i,k}_{\bar{m}} := \frac{1}{\Delta t}\left(\bar{m}_i^{k+1} - \bar{m}_i^{k}\right) - \left(u^i(n_f)_i^k + \frac{1}{2\Delta x}\left(\bar{\sigma}_{i+1}^{k} - \bar{\sigma}_{i-1}^{k}\right) - \Gamma A^0(1-\gamma^2)\bar{m}_i^k.\right)$$

and analogously $e^{i,k}_{\bar{\sigma}}$ for the second moment.

In Figure 4.6 - 4.8, the values of the previously defined error functions are shown in the discrete position-time domain with a bicubic interpolation between the discrete cells. On the abscissa there is always the discretized position domain and, on the ordinate, the discretized time interval plotted.

Notice that in these errors, two essential different kinds of errors occur. On the one hand, there is the error induced by the probabilistic nature of Monte-Carlo methods and the resulting noisy data. On the other hand, there is a truncation error caused by the discretization of the derivatives in space and time. We use always a first-order method and time and second-order method in space.

In all plots it can be observed that the error increases within time but is everywhere quite small. That the order of the error in Figure 4.6 is of two orders of magnitude smaller than in Figures 4.7 and 4.8 can be explained by the fact that in this case, the stochastic behaviour of the Keilson-Storer kernel has no influence since it cancels out by means of Lemma 4.1.

To validate the behaviour of our method, we perform experiments in which we increase the number of simulation particles. For this goal, we consider a simple example in

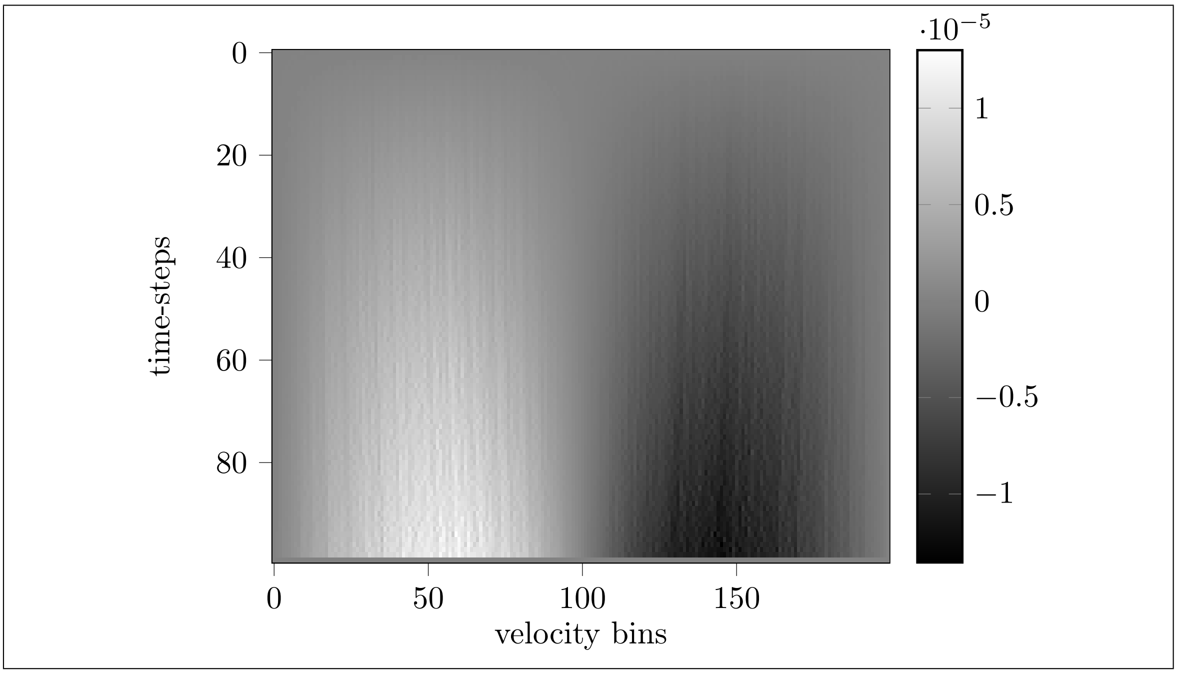


Figure 4.7. Error $e_{\bar{m}}^{i,k}$ of first moment.

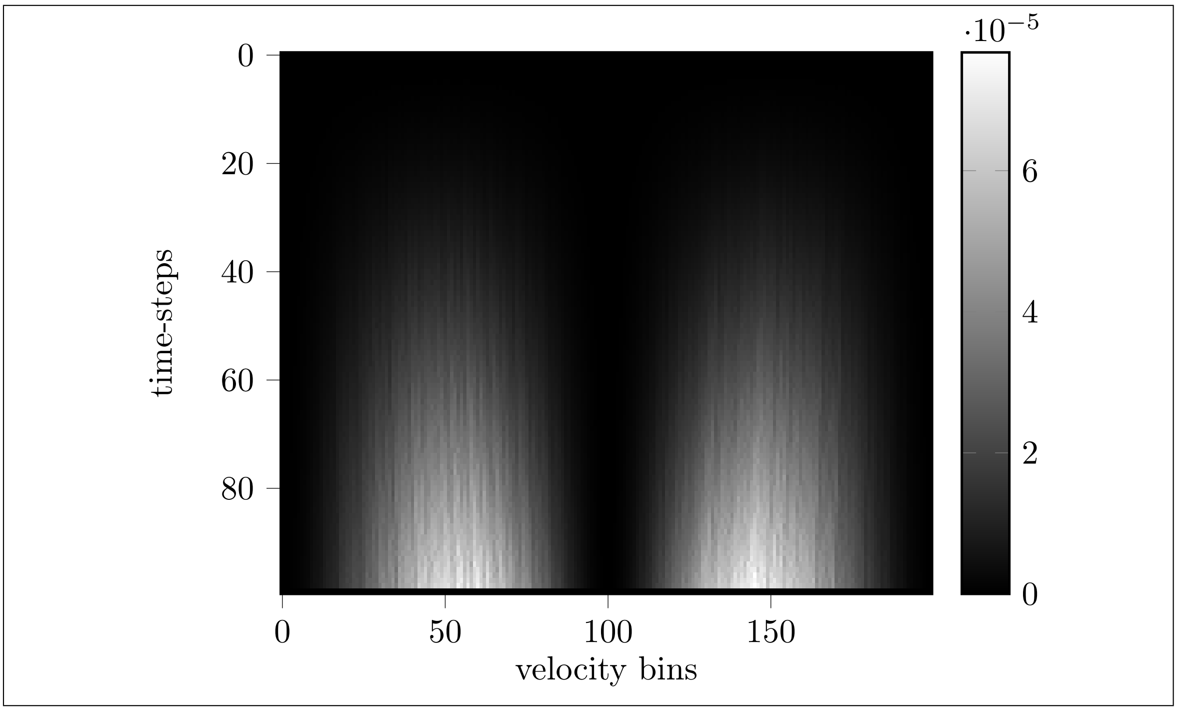
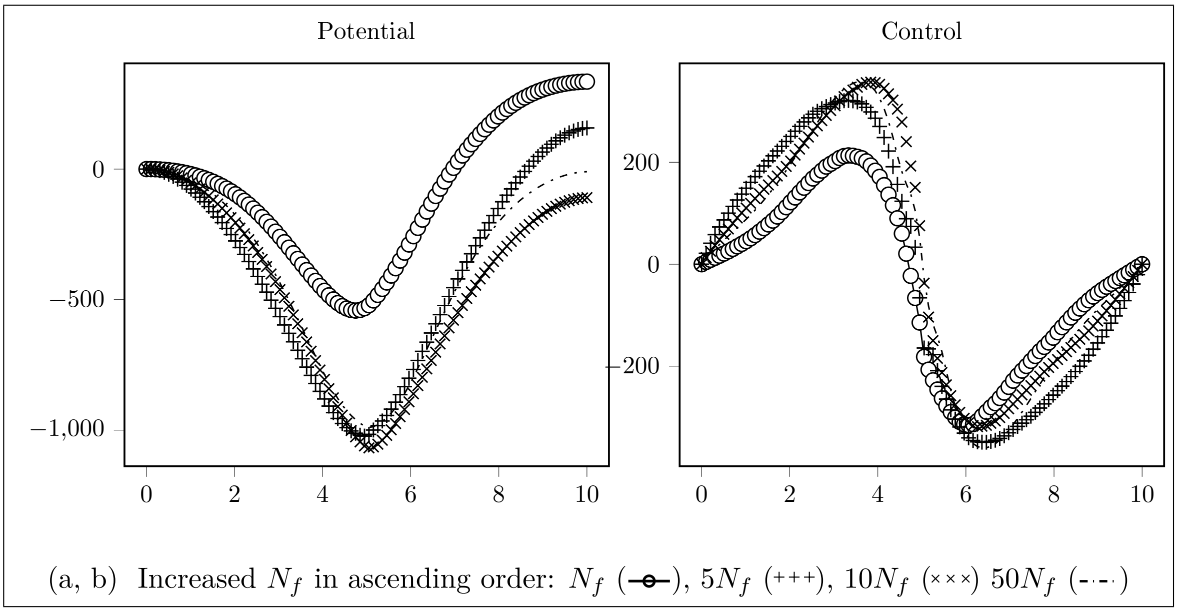


Figure 4.8. Error $e_{\bar{\sigma}}^{i,k}$ of second moment.

which the code should find a control that leads to centring the particles in the middle of the domain.



(a, b) Increased N_f in ascending order: N_f (–o–), $5N_f$ (+++), $10N_f$ (×××) $50N_f$ (-·-·)

Figure 4.9. Verification of control with respect to number of particles.

Therefore, we expect a force that is similar to the one corresponding to a single charged particle inside the domain, that is

$$u(x) \sim \frac{1}{x-L}.$$

For the purpose of the verification, we introduce the well-known notation of a potential corresponding to a force. Specifically, to a control u, we define the potential ψ_u as

$$\psi_u(x) = -\int_0^x u(\xi)\, d\xi.$$

Hence, in the present case, we expect a potential as

$$\psi_u(x) \sim -\frac{1}{(x-L)^2}.$$

However, by the choice of our space in which we search for controls, we need that the control is zero at the boundary at the domain and further continuous. Further, we expect, that our control is symmetric with respect to the position domain and has a minimum located at its centre. To perform the verification, we proceed as follows. We take the same parameters as in the next experiment. Specifically, as they are presented below in Table 4.2 and just alter the numbers of simulation particles N_f. This is due to the fact, that the accuracy of Monte Carlo methods is $\mathcal{O}(1/\sqrt{N_f})$ as a consequence of the central limit theorem; see, e.g., [**37**].

Therefore, it is a sensible approach to increase the numbers of simulation particles N_f and expect a better behaviour of the result. We started with $N_f = 10^3$ and increased the number of particles with the factor five several times. The resulting controls and corresponding potentials are plotted in Figure 4.9.

Notice that the structure of the obtained controls is close to what we expect from the reasoning above. In particular, we see that actually

$$\psi_u(x) \sim -\exp\left(-(x-L)^2\right), \qquad u(x) \sim -\psi'(x).$$

Therefore, the expected fundamental structure of the potential is recovered, specifically it has a unique minimum, is radially decreasing to this minimum and is symmetric with respect to the position domain.

Notice that the property of the potential to have a single minimum in the centre of the domain is already obtained for less particles and in the control picture, the control appears to be symmetric. However, in the potential plot it is evident that the symmetry of obtain only for higher numbers of simulation particles.

Next, we show results of numerical experiments to validate the ability of our optimal control scheme to drive the kinetic model to follow a given trajectory that corresponds to the dynamics of an harmonic oscillator in the phase space.

We test our optimization framework in a one dimensional domain $[0, L]$, $L = 10m$ in position for Argon, and we take the physical particles with mass $M = 6.63 \cdot 10^{-26} kg$; further, we combine ω_f particles in one simulation particle. The total number of simulation particles is denoted with N_f. We assume that initially the particles are normally distributed in space and velocity as follows

$$f_0 = \mathcal{N}_2\left(z_0, \Sigma_0\right), \qquad z_0 = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} 5.0 \\ 0.0 \end{pmatrix}, \qquad \Sigma_0 = \begin{pmatrix} 0.15 & 0 \\ 0 & 5.0 \end{pmatrix}.$$

We consider the optimal control problem (4.43) with $d = 1$ and the following setting. We have

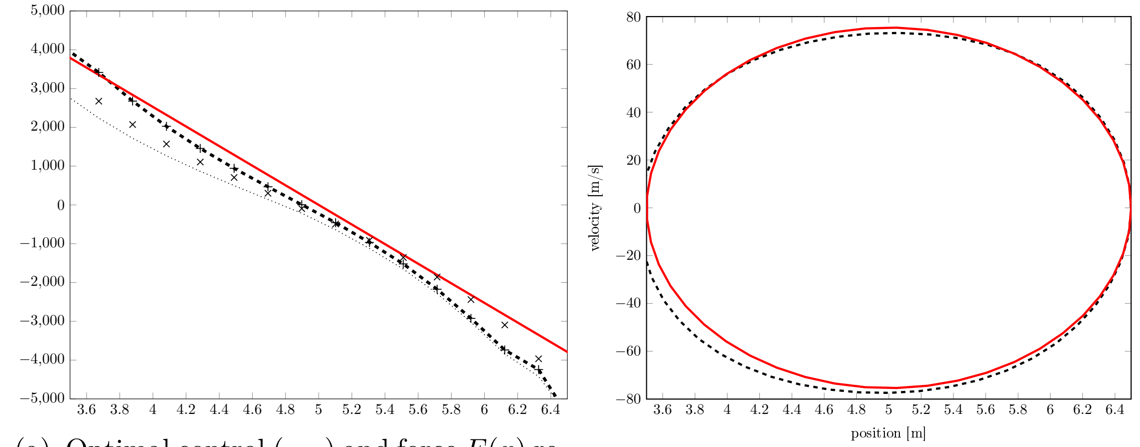
$$\begin{aligned} \zeta_D(t) &= \begin{pmatrix} 1.5\cos(\omega t) + x_0 \\ -1.5\omega\sin(\omega t) - v_0 \end{pmatrix}, & u^0(t) := (0,0)^T \\ \theta(z,t) &= -\frac{C_\theta}{2\pi\sigma_x\sigma_v}\exp\left(-\frac{|x-\xi_D(t)|^2}{2\sigma_x^2} - \frac{|v-\eta_D(t)|^2}{2\sigma_v^2}\right), \\ \varphi(z) &= 10^{-16}\theta(z,T), \qquad C_\theta = 10^{15}. \end{aligned}$$

We take $\sigma_x = 1.5$ and $\sigma_v = 30$; these parameters determine the 'width' of the potentials, that is, their effective basin of attraction. On the other hand, they enter as the variance of the distribution with which adjoint particles are created in every time-step.

In our optimization procedure, we initialize with $u^0 \equiv 0$. Further, we choose the values of the physical and numerical parameters as given in Table 4.2.

Symbol	Value	Symbol	Value
T_p $[K]$	10^3	γ [-]	0.9999
ω_f [-]	$9 \cdot 10^8$	N_f [-]	$5 \cdot 10^3$
T $[s]$	0.125	N_{frac} [-]	$2 \cdot 10^2$
N_t [-]	50	Δt $[s]$	$2.5 \cdot 10^{-3}$
$N_x \times N_v$ [-]	50×25	$v_{\max}$ $[m/s]$	10^2
ν $[\frac{s^2}{m^2}]$	10^{-6}	Δv $[m/s]$	8

Table 4.2. Physical and numerical parameters.



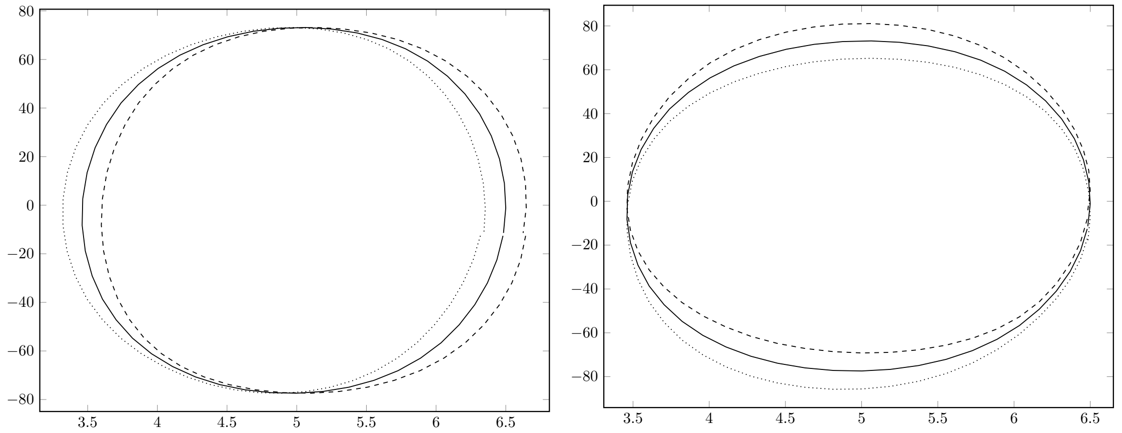
(a) Optimal control (- - -) and force $F(x)$ related to the harmonic oscillator(——), compared to two times N_f (·····), four times N_f (×××), eight times N_f (+++).

(b) Trajectory in phase space $(\langle x \rangle, \langle v \rangle)$: desired (——) and numerical (- - -)

Figure 4.10. Results of numerical experiment in the H_0^1 case (including collision). The bottom axis represents the position in both pictures. The left axis in (a) represents the value of the control, and in (b) the velocity. We use SI - units.

In Figure 4.10a (left), we depict the resulting optimal control (force) obtained with Algorithm 4.15 with our choice of N_f particles. This result is compared with those obtained with larger numbers of simulation particles, showing that the resulting optimal force does not significantly change. The axis of abscissa corresponds to the position-coordinate. For comparison, we also plot the force $F(x)$ for the harmonic oscillator whose dynamics corresponds to the desired trajectory. Specifically, this force is given by

$$F(x) = -\omega^2(x - \mu), \qquad \omega = \frac{2\pi}{T}, \qquad \mu = \frac{L}{2}.$$



(a) Trajectories of mean value $\langle x \rangle (t)$ (—), $\langle x \rangle (t) + \sigma_x(t)$ (- - -), $\langle x \rangle (t) - \sigma_x(t)$ (·····)

(b) Trajectories of mean value $\langle v \rangle (t)$ (—), $\langle v \rangle (t) + \sigma_v(t)$ (- - -), $\langle v \rangle (t) - \sigma_v(t)$ (·····)

Figure 4.11. Mean and variance of the optimal trajectory in phase space. The bottom axis represents the position and the right axis the velocity in both pictures.

In Figure 4.10b (right), we plot the evolution of the mean position and mean velocity computed as follows

$$\langle v \rangle (t) = \iint v\, f(x,v,t)\, dx\, dv, \qquad \langle x \rangle (t) = \iint x\, f(x,v,t)\, dv\, dx. \tag{4.55}$$

We see that the optimal control force obtained with our procedure effectively drives the ensemble of particles to accurately follow the given trajectory.

In Figure 4.11, we plot the mean values of position and velocity together with these values plus and minus the corresponding variances in phase space, σ_x and σ_v, respectively. These variances are calculated as follows

$$\sigma_x^2(t) = \left\langle x^2 \right\rangle (t) - \langle x \rangle^2 (t), \qquad \sigma_v^2(t) = \left\langle v^2 \right\rangle (t) - \langle v \rangle^2 (t).$$

The experiments in this Chapter no show that MOCOKI is able to calculate an optimal control to our ensemble optimal control problem. With these experiments our investigation of ensemble optimal control problems governed by kinetic models with linear collision kernel is finished. Using our approach, it is possible to define control mechanisms within the collision kernel or the external force and calculate an optimal control, such that the mean of the ensemble of particles follows a desired trajectory.

Chapter 5

Towards the control of non-linear kinetic models

In the previous chapters, we have investigated optimal control problems governed by kinetic equations with linear collision kernels. This chapter gives an outline of preliminary work on non-linear collision models. The physical interpretation of such models is that of a gas consisting of particles of a single species in which only binary collisions between particles occur. This is the case of dilute gases [**43**]. We fix the dimension $d = 3$ and consider a bounded domain $\Omega \subset \mathbb{R}^3$ with at least piece-wise smooth boundary $\partial\Omega$. For our theoretical investigations, we need in fact a boundary $\partial\Omega \in C^2$. In this case with $n(x)$, we denote the outward unit normal vector of $\partial\Omega$ at $x \in \partial\Omega$. To ease the implementation for our numerical experiments, we perform them in domains having only piece-wise smooth boundary.

The collisions are assumed to be fully elastic to ensure conservation of linear momentum and energy. In this case, we have the following relation between the post-collision velocities v', w' and pre-collision velocities v, w

$$v' = v - \mathfrak{n}\,(\mathfrak{n}, v - w)\,, \qquad\qquad w' = w - \mathfrak{n}\,(\mathfrak{n}, v - w)\,,$$

where $\mathfrak{n}$ is the unit vector directed along the line joining the centres of the two particles. We introduce the following standard notation to shorten the presentation of the forthcoming equations. For fixed x and t, we define

$$f = f(x, v, t), \qquad\qquad f_* = f(x, w, t),$$

and analogous f' and f'_* in correspondence to v' and w', respectively. The Boltzmann collision term C_B then reads

$$C_B[f, f] = \int_{\mathbb{R}^3} \int_{\mathbb{S}^2} c(|v - w|, \vartheta)(f' f'_* - f f_*)\, d\vartheta\, dw. \tag{5.1}$$

In (5.1), ϑ denotes a unit vector, $\mathbb{S}^2$ the two-dimensional unit sphere, and c the collision kernel for the scattering from (v, w) to (v', w'), see [**110**]. We further assume that the so-called detailed balance holds, this is,

$$c(|v - w|, \vartheta) = c(|v' - w'|, \vartheta).$$

5.1. Formulation of the non-linear kinetic optimal control problem

We consider the control mechanism to be within the external force $\mathcal{F}$ in the following way. For a differentiable vector field $a_0 : \mathbb{R}^3 \to \mathbb{R}^3$ and parameters $a_1, a_2 \in \mathbb{R}$, we define the force as

$$\mathcal{F}(x, v; u) := \Big(a_0(x) + a_1 u_1(x) + a_2\, v \times u_2(x)\Big), \tag{5.2}$$

where × denotes the standard cross product for two vectors in three dimensions. In (5.2), we consider the functions $u_\iota(x) : \mathbb{R}^3 \to \mathbb{R}^3$, $\iota = 1, 2$ as the controls. Notice that by this structure, the control u_1 acts like an electric force and u_2 acts like a magnetic force. We define $u : \mathbb{R}^6 \to \mathbb{R}^6$, $u(x) = (u_1(x), u_2(x))$.

Using $\mathcal{F}$, we introduce the controlled free streaming operator as follows,

$$L_u \;:=\; v \cdot \nabla_x \;+\; \mathcal{F}(x, v; u) \cdot \nabla_v \;, \tag{5.3}$$

where we set the mass of the particles to $M = 1$. For the control space, we choose $U = H_0^2(\Omega) \times H_0^2(\Omega)$. This choice is motivated by our numerical experiments that show almost smooth controls in this case. This property could not be obtained considering controls in $H_0^1(\Omega)$. The theoretical explanation for this may be that the control is at least Hölder continuous with Hölder index 1/2 for $u \in U$. This follows from Sobolev embeddings in three dimensions [**1, 66**]. We endow the control space U with its natural norm

$$\|u\|_U = \left(\sum_{\iota=1}^{2} \int_\Omega |u_\iota(x)|^2\, dx + \int_\Omega |\nabla u_\iota(x)|^2\, dx + \int_\Omega |\nabla^2 u_\iota(x)|^2\, dx\right)^{1/2}.$$

Now, we can write a non-linear kinetic equation of the form

$$\partial_t f(x, v, t) \;+\; L_u f(x, v, t) = C_B[f, f](x, v, t),$$

for $(x, v, t) \in \Omega \times \mathbb{R}^3 \times [0, T]$. On the boundary $\partial\Omega$ of the domain, we consider specular reflection at inflow. We can express this fact using the operator $Rv = v - 2(n \cdot v)n$ as

$$f(x, v, t) = f(x, Rv, t) \qquad \text{on} \qquad \partial\Omega \times \mathbb{R}^3_< \times [0, T].$$

Moreover, we define an initial condition $f_0(x,v) \in H_k^m(\Omega \times \mathbb{R}^3)$. Now, we can formulate the non-linear kinetic model as follows

$$(5.4)\qquad \begin{cases} \partial_t f(x,v,t) \,+\, L_u f(x,v,t) \,=\, C_B[f,f](x,v,t) & \text{in} \quad \Omega \times \mathbb{R}^3 \times (0,T] \\ f(x,v,t) = f(x,Rv,t) & \text{on} \quad \partial\Omega \times \mathbb{R}^3_< \times (0,T] \\ f(x,v,0) = f_0(x,v) & \text{on} \quad \Omega \times \mathbb{R}^3. \end{cases}$$

Notice that very recently, a similar problem has been discussed in detail in [**87**]. Let us introduce the control-to-state map G that maps a given control $u \in U$ for a fixed initial condition f_0 to the solution of (5.4):

$$G : U \longrightarrow L_T^\infty\left(L^2(\Omega \times \mathbb{R}^3)\right), \qquad\qquad u \mapsto f = G(u).$$

We assume that G is well-defined and continuous, as underpinned by the results of [**87**, Theorem 1.1]. In this paper, the authors derive an existence and uniqueness result under further regularity assumptions on the boundary, the integrability of the initial condition, and regularity and smallness assumption on the force and the initial condition. Further, existence and uniqueness results are presented for the unbounded space or periodic boundary condition in [**75**, Theorem 3.2] using a cut-off and in [**4**, Theorem 1.1], [**63**, Theorem 1.1] without such an assumption. In the last two references, it is assumed that the initial condition is close enough in a certain sense to the equilibrium solution.

To formulate the optimal control problem, we consider the following objective. We define $\theta(\cdot,t)$ and $\varphi(\cdot)$ as negative Gaussian functions in phase-space and the functional

$$(5.5)\qquad J(f,u) := \int_0^T \int_{\Omega\times\mathbb{R}^3} \theta(z,t)\, f(z,t)\, dz\, dt + \int_{\Omega\times\mathbb{R}^3} \varphi(z)\, f(z,T)\, dz + \frac{\nu}{2}\, \|u\|_U^2\,.$$

The parameter $\nu > 0$ describes the weight of the cost of the control.

With this preparation, we can state our optimal control problem as

$$(5.6)\qquad \begin{aligned} &\min\ J(f,u) \\ &\text{s.t.} \begin{cases} \partial_t f(x,v,t) \,+\, L_u f(x,v,t) \,=\, C_B[f,f](x,v,t) & \text{in} \quad \Omega \times \mathbb{R}^3 \times (0,T] \\ f(x,v,t) = f(x,Rv,t) & \text{on} \quad \partial\Omega \times \mathbb{R}^3_< \times (0,T] \\ f(x,v,0) = f_0(x,v) & \text{on} \quad \Omega \times \mathbb{R}^3 \end{cases} \\ &u \in U. \end{aligned}$$

5.2. Non-linear kinetic optimality system

To characterize optimal controls to (5.6), we derive the corresponding optimality system. This system consists of the kinetic model (5.4), an adjoint kinetic model and the reduced gradient. The latter one is formulated using the solutions of the former two models. Following the Lagrange approach, we define the Lagrange functional with the Lagrange multipliers q, q_0, q_σ as follows

$$\begin{aligned}\mathcal{L}(f,u,q,q_0,q_\sigma) \coloneqq J(f,u) + \int_0^T\!\!\int_{\Omega\times\mathbb{R}^3} \Big(\partial_t f(z,t) \,+\, L_u f(z,t) \,-\, C_B[f,f](z,t)\Big)\, q(z,t)\, dz\, dt \\ + \int_{\Omega\times\mathbb{R}^3} \left(f(z,0) - f_0(z)\right) q_0(z)\, dz \\ + \int_{L^2(\partial\Omega\times\mathbb{R}^3_<\times[0,T])} \left(f(x,v,t) - f(x,Rv,t)\right) q_\sigma(x,v,t)\, d\sigma\, dv\, dt,\end{aligned}$$

To derive the adjoint equation, we have to consider the derivative of $\mathcal{L}$ with respect to f. The most challenging part in this calculation is to obtain the derivative of the collision kernel C_B that leads to an adjoint collision kernel C_B^*. Let δf be an arbitrary, small variation of f, such that $f + \alpha\,\delta f \in L_T^\infty H_k^m$ for small $\alpha > 0$ and $\delta f_{|t=0} = 0$. The derivative of the collision part is then given as the limit for $\alpha \to 0$ of the following expression divided by α

$$\begin{aligned}\int_{\mathbb{R}^3} dv\, q(x,v,t) \int_{\mathbb{R}^3} dw \int_{\mathbb{S}^2} d\vartheta\, c(|v-w|,\vartheta)((f' + \alpha\,\delta f')(f'_* + \alpha\,\delta f'_*) - (f + \alpha\,\delta f)(f_* + \alpha\,\delta f_*)) - \\ \int_{\mathbb{R}^3} dv\, q(x,v,t) \int_{\mathbb{R}^3} dw \int_{\mathbb{S}^2} d\vartheta\, c(|v-w|,\vartheta)(f'f'_* - ff_*).\end{aligned}$$

This expression is after a division by α and passing to the limit $\alpha \to 0$ equivalent to

$$\int_{\mathbb{R}^3} dv\, q(x,v,t) \int_{\mathbb{R}^3} dw \int_{\mathbb{S}^2} d\vartheta\, c(|v-w|,\vartheta)\Big(f'\,\delta f'_* + f'_*\,\delta f' - f\,\delta f_* - f_*\,\delta f\Big). \tag{5.7}$$

For the collision kernel, it holds by the symmetry of absolute value and detailed balance that

$$c(|v-w|,\vartheta) = c(|v'-w'|,\vartheta) = c(|w-v|,\vartheta) = c(|w'-v'|,\vartheta).$$

Therefore, renaming w, v, w', v' we see that (5.7) equals

$$\int_{\mathbb{R}^3} dv\, \delta f \int_{\mathbb{R}^3} dw \int_{\mathbb{S}^2} d\vartheta\, c(|v-w|,\vartheta) f_*(q' + q'_* - q - q_*).$$

Since δf is considered to be arbitrary, we define the adjoint collision kernel as

$$C_B^*[q](v,t) = \int_{\mathbb{R}^3}\int_{\mathbb{S}^2} c(|v-w|,\vartheta) f_*(q' + q'_* - q - q_*)\, d\vartheta\, dw. \tag{5.8}$$

Hence, $C_B^*[q]$ is linear in q and depending on f_*. Using a cut-off of the interaction potential, it is possible to rewrite the equation (5.8) and split the collision operator in a gain and a loss term, see the result of Grad, specifically (3.15) and (3.16) in [**78**]. Notice that an equivalent adjoint collision kernel can also be derived using the linearized Boltzmann equation [**36**]. For further information and the connection of the linear and linearized Boltzmann equations see [**40**] and [**64**].

Recall that the adjoint equation does not need to have a physical interpretation and its solely purpose is to be used in the calculation of the gradient. Hence, a sufficiently good approximation of it is enough for our purpose. In particular, assuming that we can approximate f_* by a Maxwell distribution $f^{eq}(w)$ we can define our adjoint collision kernel as follows

$$A_B^*(w,v) = f^{eq}(w) \int_{\mathbb{S}^2} c(|v-w|,\vartheta)\, d\vartheta. \tag{5.9}$$

Such approximation is in particular sensible in the regime of dense gases [**44**, Section 15.51]. This may be not the case for the forward equation, however, recall that the main purpose of the adjoint equation is to constitute the gradient. Using (5.9), the approximated adjoint collision term can be written in the form

$$\begin{aligned} C_B^*[q] &= \int_{\mathbb{R}^3} A_B^*(w,v) q(w)\, dw - q(v) \int_{\mathbb{R}^3} A_B^*(v,w)\, dw \\ &= \int_{\mathbb{R}^3} A_B^*(w,v) q(w)\, dw - q(v) \frac{1}{\tau^*(v)}. \end{aligned}$$

The quantity $\frac{1}{\tau^*(v)}$ is the adjoint collision frequency given by

$$\frac{1}{\tau^*(v)} = \int_{\mathbb{R}^3} A_B^*(w,v)\, dw.$$

This collision frequency depends on the velocity. However, although the adjoint equation has a linear collision kernel, it has not the structure of a standard linear approximation, as for example the BGK approximation. In fact, the dependence of the frequency on the velocity leads to theoretical and computational issues with deterministic models, which have been recently investigated [**81**].

Bringing together all the pieces for the adjoint model, we obtain the following

$$\begin{cases} -\partial_t q(x,v,t) + L_u^* q(x,v,t) = C_B^*[q](x,v,t) - \theta(x,v,t) & \text{on } \Omega \times \mathbb{R}^3 \times [0,T) \\ q(x,v,t) = q(x,Rv,t) & \text{in } \Omega \times \mathbb{R}^3_> \times [0,T) \\ q(x,v,T) = -\varphi(x,v) & \text{in } \Omega \times \mathbb{R}^3, \end{cases} \tag{5.10}$$

where we introduce the adjoint free-streaming operator

$$L_u^* = -L_u.$$

Observe that in (5.10) there is no constant source term proportional to q appearing. This is due to the symmetry of the collision kernel and hence no reformulation is needed to interpret the adjoint model as a kinetic model.

The next step is to compute the derivative of $\mathcal{L}$ with respect to the control. Recall that the control is included within the force $\mathcal{F}$. The derivatives of $\mathcal{F}$ are given by

$$\partial_{u_1}\mathcal{F}(x,v;u) = a_1 I_3, \qquad \partial_{u_2}\mathcal{F}(x,v;u) = a_2 \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix},$$

where I_3 is the three-dimensional identity matrix. We define the differential operator

$$\mathcal{D}\phi = \phi - \Delta\phi + \Delta^2\phi,$$

where Δ denotes the Laplace operator with respect to position.

With this definition, the L^2 gradient of our optimal control problem (5.6) is given by

$$\begin{aligned} \nabla_{u_\iota} J_r(u)|_{L^2}(x) &= \nu\, \mathcal{D}u_\iota(x) + \int_0^T\!\!\int_{\mathbb{R}^3} q(x,v,t)\,\partial_{u_\iota}\mathcal{F}(x,v;u)\cdot\nabla_v f(x,v,t)\,dv\,dt \\ x \in \Omega, &\qquad \iota = 1,2. \end{aligned} \tag{5.11}$$

Notice that we require control field to be an element of $H_0^2 \times H_0^2$. Thus, we have to derive a formula for $\nabla_{u_\iota} J_r(u)|_{H^2}(x)$. To shorten the presentation, we define $\psi(x) = \nabla_u J_r(u)|_{H^2}(x)$, where $\nabla_u J_r(u)|_{H^2} = (\nabla_{u_1} J_r(u)|_{H^2}, \nabla_{u_2} J_r(u)|_{H^2})$. Recall that the Taylor series expansion in a Hilbert space X must be identical term by term regardless of the choice of X; see [**28**]. In other words, we need to calculate the Riesz representant of the derivative of the reduced functional in the correct Hilbert space. Therefore, we have

$$(\psi, \delta u)_{H^2} = (\nabla_u J_r(u)|_{L^2}, \delta u)_{L^2}\,.$$

Using the definition of the H_0^2 inner product $(u,v)_{H^2} = (u,v)_{L^2} + (\nabla u, \nabla v)_{L^2} + (\Delta u, \Delta v)_{L^2}$ for functions that have compact support, we obtain the relation

$$\int_\Omega \Big(\psi(x)\,\delta u(x) + \nabla\psi(x)\cdot\nabla\delta u(x) + \Delta\psi(x)\,\Delta\delta u(x)\Big)\,dx = \int_\Omega \nabla_u J_r(u)|_{L^2}(x)\,\delta u(x)\,dx,$$

which must hold for all test functions $\delta u \in U$. To shift the derivative of ψ in the second and third term, we use Green's first and second identity in three dimensions. With the assumption that ψ and its directional derivative with respect to n are zero at the boundary $\partial\Omega$, we obtain the following boundary value problem for the H^2 gradient

$$\begin{cases} \psi(x) - \Delta\,\psi(x) + \Delta^2\,\psi(x) \;=\; \nabla_u J_r(u)|_{L^2}(x) & \text{in} \quad \Omega \\ \psi(x) = 0, \qquad \partial_n\psi(x) = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{5.12}$$

With ∂_n, we denote the derivative with respect to the outward unit normal n. We remark that the assumption that the control is zero at the boundary can be replaced by the hypothesis that the value of the control at the boundary is any given value.

The structure of (5.12) will not be altered by changing the value of the control at the boundary.

Summarizing, we have the non-linear kinetic optimality system in $\Omega \times \mathbb{R}^3 \times [0, T]$

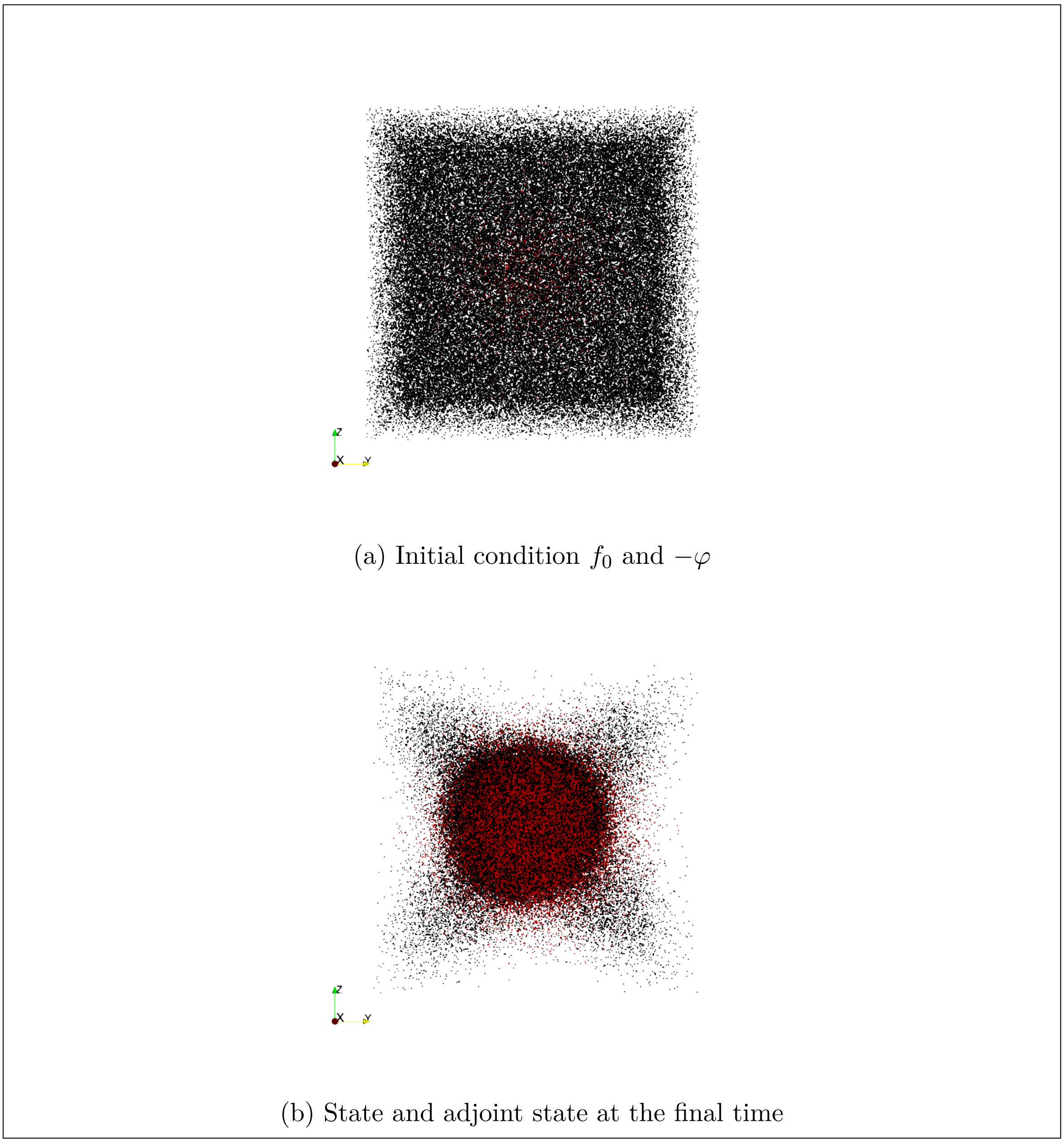
$$
\begin{aligned}
&\partial_t f(x,v,t) + L_u f(z,t) = C_B[f,f](x,v,t),\\
&f(x,v,0) = f_0(x,v),\\
&f|_{\partial\Omega\times\mathbb{R}^3_<} = f(x, v - 2n(n\cdot v), t),\\
\\
&-\partial_t q(z,t) + L_u^* q(z,t) = C_B^*[q](z,t) - \theta(z,t),\\
&q(z,T) = -\varphi(z),\\
&q|_{\partial\Omega\times\mathbb{R}^3_>} = q(x, v - 2n(n\cdot v), t),\\
\\
&\mathcal{D}\psi_\iota(x) = \nu\mathcal{D}u_\iota(x) + \int_0^T \int_{\mathbb{R}^3} q(x,v,t)\,(\partial_{u_\iota}\mathcal{F}(x,v,t;u)\nabla_v f(x,v,t))\,dv\,dt,\\
&\psi_{\iota|\partial\Omega} = 0, \qquad \partial_n\psi_{\iota|\partial\Omega} = 0, \qquad \iota = 1,2.
\end{aligned}
\tag{5.13}
$$

5.3. Results of numerical experiments

For solving the kinetic model and the adjoint kinetic model, we develop a new module within the `C++` codes of our industrial partner. Notice that (5.12) is a vector problem for the two components of the gradient $\nabla_u J_r(u)|_{H^1}(x)$. We approximate the differential operator $\mathcal{D}$ in this problem by standard finite difference approximation for the second and fourth order derivative. Specifically, we use second-order central finite difference approximation for the derivatives; for the coefficients of the discrete derivatives see [**69**]. The problem (5.12) then results in a block-penta-diagonal system. The solution of this system is obtained efficiently by an adapted Thomas method; see [**15**]. In principle, also higher order of approximations of derivatives can be used, which then will need information of more cells near the cell under consideration. This will lead to more complicated and less sparse matrices. For approximating the integral, we use the second-order trapezoidal rule.

Now, we present results of numerical experiments in order to validate the ability of our control framework to calculate solutions of our optimization problem (5.13) in the six-dimensional phase-space. As $\Omega \subseteq \mathbb{R}^3$ we consider a cube and assume specular reflection as the boundary condition. We assume that the control acts on the particles as electric or magnetic force. No further interactions between the particles are considered, in particular no electrostatic forces.

In the first test-case, we choose $a_0 \equiv 0$, $a_1 = 1$ and $a_2 = 0$ in the control mechanism (5.2). The initial distribution is taken to be uniform in position and in thermal



(a) Initial condition f_0 and $-\varphi$

(b) State and adjoint state at the final time

Figure 5.1. Distribution of forward (black) and adjoint (red) particles in the domain Ω. The axis x_i, $i = 1, 2, 3$ are labelled in their physical more common notation x, y, z.

equilibrium in velocity. The desired state is a Gaussian distribution in both position and velocity with the mean at position being the centre of the box. This desired state is encoded in θ and φ.

In Figure 5.1, it is shown that our calculated control accomplishes the task to drive the particle distribution close to the desired state. Specifically, the model particles are gathered in the centre of the domain Ω. Notice that the edges in the Figure 5.1b are due to the fact that we start with homogeneously distributed particles in the position domain, but the desired state is Gaussian, that is a ball with respect to

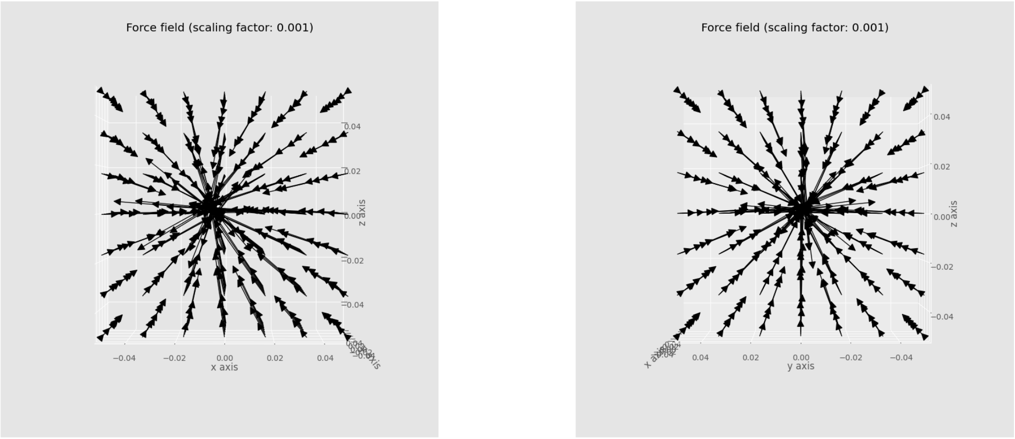


Figure 5.2. Calculated force from different perspectives.

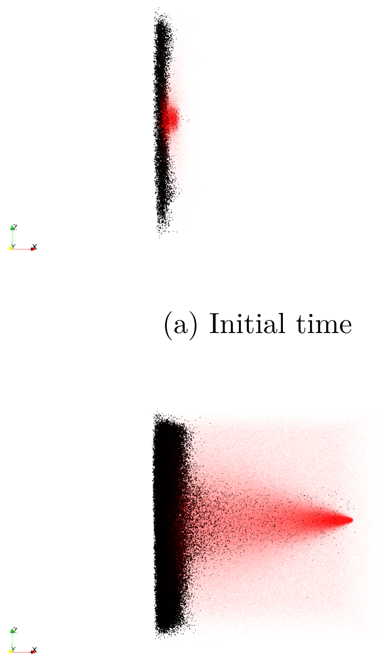
position. Therefore, by the expected spherical symmetry of the control, particles at the edges of the box need more time to be pushed in the centre. Thus, the behaviour shown in Figure 5.1b is to be expected.

In Figure 5.2, the control is plotted from different perspectives. Notice that the control is not perfectly symmetric which we would expect from a control that fits perfectly to the desired state. However, while increasing the number of the simulation particles, we see convergence of the control to an entirely symmetric one in our numerical experiments.

In the next test case, we consider $a_1 = 0$ and $a_2 = 1$ in (5.2). We consider an inflow from one side of the boundary and aim to decrease the area occupied by particles during their movement to the other side of the box.

In Figure 5.3, the state corresponding to the calculated control is shown at different time-steps. Specifically at the initial and the terminal time. Notice that the area in which particles exist is clearly decreased compared to the area they cover at the inflow.

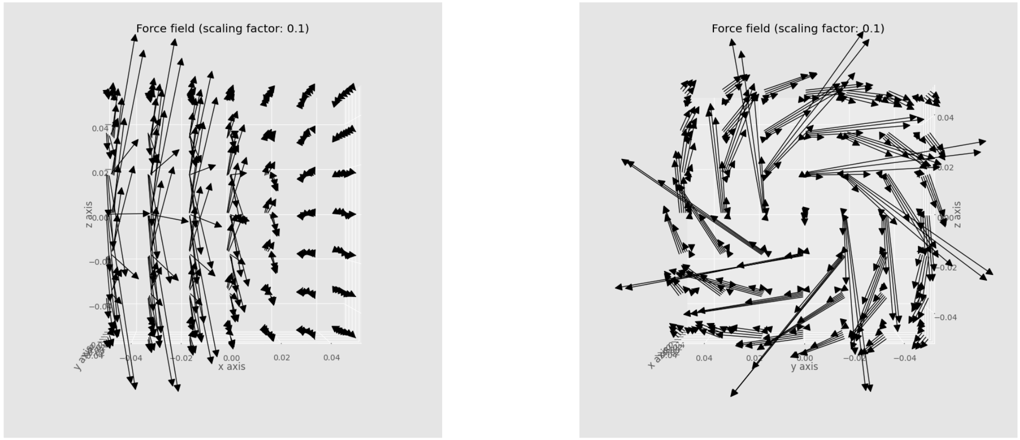
In Figure 5.4, the calculated control in the magnetic test-case is plotted from two perspectives. In Figure 5.4a the control is shown from the same perspective as the plots in Figure 5.3. Notice that the intensity $|u(x)|$ is decreasing in x_1-direction. This due to the fact that the number density is higher for smaller values of x_1 and hence the gradient for such values for x_1 is greater in absolute value. In Figure 5.4b the control is plotted with the x_2 axis at the bottom. In this picture, it is observable that the calculated control is close to a divergence-free function.



(a) Initial time

(b) Final time

Figure 5.3. Particle distributions at different time-steps.



(a) View x_1-Axis

(b) View x_2-Axis

Figure 5.4. Calculated magnetic force.

It is desirable to have a divergence-free function in view of the interpretation of the control as an external magnetic force. Since no magnetic monopoles exists, the divergence of the control should vanish. We can include this in our framework, by applying a projection in the update of our control. This is achievable due to the fact that is possible to decompose vector-fields in a divergence free and a rotation-free component [**20**]. This decomposition is the so-called Helmholtz-decomposition. Further, several algorithms have been developed to calculate such a decomposition in different dimensions; see, e.g., [**2, 80**]. Another possibility is to include the norm of the divergence as additional term in the cost functional.

5.4. Outlook

We finish this work with some final remarks and some further outlooks. In the present work, only models that resemble single species gas were considered and it was assumed that the particles do only interact via collisions. In particular, no self-induced interaction forces were considered. However, these physical assumptions can be discarded. Specifically, it is possible to investigate multi-species systems, in particular plasma including inter-particle forces. In this case, a model equation must be stated for each of the species and a term for each species will occur in the definition of the objective. Moreover, for each species an adjoint equation must be derived. However, notice that in the assembling of the gradient only the species enter on which the control is applied directly. Hence, depending on the specific control mechanism, it may be not required to solve all adjoint equations.

Further, also inter-species interactions beyond collisions could be considered. For example, electro-static forces that are modelled with Poisson's equation. Since the adjoint equation of each species is the linearization of the model equation with respect to this species, a linear correlation of the adjoint and the model variable is to be expected.

Similar to the inclusion of electro-static forces, one may consider additional Maxwell's equations describing electro-magnetic fields. In particular, if the control should be interpreted as an external magnetic force, it must fulfil the corresponding equations. In particular, it must be divergence free. These and similar restrictions can be taken into account using a projected optimization algorithm.

Until now, it is also assumed that the control acts on the whole position domain. However, there are two different possibilities to reject this assumption. On the one hand, the domain in which the control acts may be restricted using a the characteristic function smoothed by mollifiers on the domain. Outside this domain the control is set to zero. On the other hand, it is possible to define a control mechanism located at the boundary of the domain and calculate the resulting force inside with appropriate equations and numerical solvers. In this case, the control should consist of the value at the boundary and the directional derivative with respect to the unit normal vector of the domain. The advantage in such a setting is that a physical device that realizes the control mechanism may be easier obtained than in the case where the control itself is defined in the whole domain.

Moreover, one may include erosion of other chemical reactions between the particles or between particles and the physical boundary of the domain. This will lead to more realistic models. Further, using our ensemble control problems, it is possible to minimize various macroscopic properties of the density function using the structure of the objective. This is realizable by virtue of the definition of the tracking and terminal cost part in the functional as weighted mean values. Therefore, all physical quantities that can be expressed using expected values respectively moments of density functions can in principle be considered in the objective.

Appendix A

Code documentation

A.1. RESOLVE: Optimal control problems governed by the Liouville equation

The code used in the Chapters 2 and 3 is called RESOLVE. Please find on the attached Compact Disc (CD) the codes for the three experiments presented in this thesis. We now shortly explain the structure of the program.

A.1.1. Structure of the program

The code is a standalone MATLAB program that requires no further MATLAB packages. It is written using MATLAB R2020b.

The auxiliary directory of the code is structured in four categories

- **Better control**: contains several files used in the update process as projections and a line-search algorithm
- **Kurganov-Tadmor**: contains subroutines executing the SSPRK2-KT scheme
- **Semi-smooth Newton**: contains files for the calculation for the next control using a semi-smooth Krylow-Newton scheme; in particular, a file for calculating the reduced Hessian, and solving the arising linear equations and dealing with the non-smoothness if the L^1 cost is included in the functional
- **Strang splitting**: contains subroutines executing the KTS scheme

A.1.2. Using the program

The main file of the program is called `DRIVER_OptLiouv_Brockett.m` and represents Algorithm 3.2. In this file, the initial guess of the control is loaded, if the file `OCP_Liouville.mat` with the correct structure can be found in the directory. Otherwise, the default value for the initial guess is the function that is constant zero everywhere.

All the parameters for discretization and optimization are defined in the file `globalParameters.m`. In particular, the mesh in time and space is defined in this file. The parameters will be accessed during the program where needed.

The initial condition for the density function should be defined in the file `initialConditions_Brockett.m`. Recall that it should be zero near the boundary. In this file, also the potentials θ and φ that are important for the calculation of the objective and solving the adjoint problem are defined as global variables.

The objective is implemented in the file `objectiveJ_Brockett.m`. It makes use of the potentials defined in the `initialConditions_Brockett.m` file.

A.1.3. Output and post-processing

The program terminates if the difference between the controls of two consecutive optimization iterations is smaller than a given threshold or the maximal iteration depth is reached. After a successful iteration of the algorithm, the current control, the history of the value of the functional and the norm of the gradient as well as the solutions to the forward and backward problem are saved in a file called `OCP_Liouville.mat`. In particular, the obtained control and corresponding states are saved after the termination of the program. With this file, it is possible to restart the program with the values of the last iteration if needed.

During the execution of the code there are several output-figures displaying the current state of the optimization. In particular, the current value of the functional, the norm of the gradient and the current control are plotted. These figures are saved in the `pictures/` folder. Besides this, some less important figures as the process of the H^1 projection are shown.

A.2. MOCOKI: A Monte Carlo approach for optimal control in the force of a linear kinetic model

The MOCOKI code solves optimal control problems governed by linear kinetic equations including external forces and a collision term in a Monte Carlo framework. Please find on the attached CD the sources files for MOCOKI.

A.2.1. Dependencies and required libraries

The code was optimized for Ubuntu 18.04 LTS. Before downloading the dependencies, make sure that Ubuntu is up-to-date using

```
sudo apt-get update
```

and

```
sudo apt-get upgrade.
```

Before compiling the code the following dependencies and libraries must be installed: **openMP** (for parallelizing) that can be installed using

```
sudo apt install libomp-dev
```

and **cmake** that can be installed using

```
sudo apt install cmake.
```

For optional postprocessing `python3` should be installed including the packages `argparse`, `pyplot` from `matplotlib`, `numpy`, `math`, `pandas`. These packages can be installed using the following commands

```
sudo apt install python3-pip -y
pip3 install matplotlib pandas numpy
```

A.2.2. Problem specifications

In the file `globalparameters.h` it is possible to specify the parameters used in the code. View the comments in the file to get information about the purpose of each parameter. The file `optimization/optimization_algorithms.cpp` is the core of the optimization scheme and mirrors Algorithm 4.15; in this file the initial condition of the kinetic model and the adjoint kinetic model can be specified.

A.2.3. Structure of the code

The code is structured in four categories:

- **auxiliary**: contains auxiliary subroutines like generating of probability density functions (pdf) and controller for input/output
- **mcc**: contains methods for solving the linear kinetic and adjoint linear kinetic problem
- **monitoring**: contains methods for keeping track of numbers important for optimization like value of functional

- **optimization**: core of optimization methods; contains NCG subroutines and armijo-linesearch as well as functions providing the value of the functional and the building of the gradient
- **post__processing**: contains python files for visualizing the results of the simulation

A.2.4. Compiling and running the program

After specifying the parameters, it is possible to compile the code and start the program with the following commands inside the MOCOKI folder, that contains all the source files.

```
cd build-MOCOKI
cmake ../
make
./MOCOKI
```

The output files will be written in the folder output-MOCOKI.

A.2.5. Post-processing

There is a python file to produce pictures of the results of the MOCOKI code. For this purpose, change to directory post-processing and execute

```
python3 solution_kinetic_model.py ../output-MOCOKI/
```

The resulting plots show: the desired and obtained trajectories in phase space, the corresponding optimal control and the optimization history of the objective functional. These figures are saved in figures-MOCOKI, in the file solution_kinetic_model.png, together with the control's profile stored in the ascii file calculated_control.txt .

References

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, vol. 140 of Pure and Applied Mathematics (Amsterdam), Elsevier/Academic Press, second ed., 2003.

[2] E. Ahusborde, M. Azaïez, J.-P. Caltagirone, M. Gerritsma, and A. Lemoine, *Discrete Hodge Helmholtz decomposition*, in Twelfth International Conference Zaragoza-Pau on Mathematics, vol. 39 of Monogr. Mat. García Galdeano, Prensas Univ. Zaragoza, Zaragoza, 2014, pp. 1–10.

[3] V. Akçelik, G. Biros, O. Ghattas, J. Hill, D. Keyes, and B. van Bloemen Waanders, *Parallel algorithms for pde-constrained optimization*, in Parallel Processing for Scientific Computing, M. A. Heroux, P. Raghavan, and H. D. Simon, eds., Society for Industrial and Applied Mathematics, 2006, ch. 16, pp. 291–322.

[4] R. Alexandre, Y. Morimoto, S. Ukai, C.-J. Xu, and T. Yang, *Global existence and full regularity of the Boltzmann equation without angular cutoff*, Comm. Math. Phys., 304 (2011), pp. 513–581.

[5] R. Alonso, *Boltzmann-type equations and their applications*, Publicações Matemáticas do IMPA. [IMPA Mathematical Publications], Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 2015.

[6] L. Ambrosio, *Transport equation and Cauchy problem for BV vector fields*, Invent. Math., 158 (2004), pp. 227–260.

[7] L. Ambrosio and G. Crippa, *Existence, uniqueness, stability and differentiability properties of the flow associated to weakly differentiable vector fields*, in Transport equations and multi-D hyperbolic conservation laws, vol. 5 of Lect. Notes Unione Mat. Ital., Springer, Berlin, 2008, pp. 3–57.

[8] ——, *Continuity equations and ODE flows with non-smooth velocity*, Proc. Roy. Soc. Edinburgh Sect. A, 144 (2014), pp. 1191–1244.

[9] N. Ayi, *From Newton's law to the linear Boltzmann equation without cut-off*, Comm. Math. Phys., 350 (2017), pp. 1219–1274.

[10] H. Bahouri, J.-Y. Chemin, and R. Danchin, *Fourier analysis and nonlinear partial differential equations*, vol. 343 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer, Heidelberg, 2011.

[11] V. Barbu and T. Precupanu, *Convexity and optimization in Banach spaces*, Springer Monographs in Mathematics, Springer, Dordrecht, fourth ed., 2012.

[12] J. Bartsch, A. Borzì, F. Fanelli, and S. Roy, *A theoretical investigation of Brockett's ensemble optimal control problems*, Calc. Var. Partial Differential Equations, 58 (2019), p. Paper No. 162.

[13] J. Bartsch, G. Nastasi, and A. Borzì, *Optimal control of the Keilson-Storer master equation in a Monte Carlo framework*, J. Comput. Theor. Transp., (2021).

[14] R. Beals and V. Protopopescu, *Abstract time-dependent transport equations*, J. Math. Anal. Appl., 121 (1987), pp. 370–405.

[15] K. Benkert and R. Fischer, *An efficient implementation of the thomas-algorithm for block penta-diagonal systems on vector computers*, in International Conference on Computational Science, Springer, 2007, pp. 144–151.

[16] P. Berman and V. Malinovsky, *Principles of Laser Spectroscopy and Quantum Optics*, Princeton University Press, Princeton, 2010.

[17] P. R. Berman, J. E. M. Haverkort, and J. P. Woerdman, *Collision kernels and transport coefficients*, Phys. Rev. A, 34 (1986), pp. 4647–4656.

[18] D. Bernoulli, *Hydrodynamica, sive de viribus et motibus fluidorum commentarii*, Sumptibus J.R. Dulseckeri, Argentorati, 1738.

[19] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*, Computer Science and Applied Mathematics, Academic Press, Inc., New York-London, 1982.

[20] H. Bhatia, G. Norgard, V. Pascucci, and P.-T. Bremer, *The helmholtz-hodge decomposition—a survey*, IEEE Trans. Vis. Comput. Graph., 19 (2012), pp. 1386–1404.

[21] P. L. Bhatnagar, E. P. Gross, and M. Krook, *A model for collision processes in gases. i. small amplitude processes in charged and neutral one-component systems*, Phys. Rev., 94 (1954), pp. 511–525.

[22] G. A. Bird, *Monte carlo simulation of gas flows*, Annual Review of Fluid Mechanics, 10 (1978), pp. 11–31.

[23] ———, *The DSMC Method*, Create Space, v1.2 ed., 2013.

[24] T. Bodineau, I. Gallagher, and L. Saint-Raymond, *The Brownian motion as the limit of a deterministic system of hard-spheres*, Invent. Math., 203 (2016), pp. 493–553.

[25] L. Boltzmann, *Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen*, Sitz.-Ber. Akad. Wiss. Wien (II), 66 (1872), pp. 275–370.

[26] A. Borzì, G. Ciaramella, and M. Sprengel, *Formulation and numerical solution of quantum control problems*, vol. 16 of Computational Science & Engineering, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2017.

[27] A. Borzì, K. Kunisch, and D. Y. Kwak, *Accuracy and convergence properties of the finite difference multigrid solution of an optimal control optimality system*, SIAM J. Control Optim., 41 (2002), pp. 1477–1497.

[28] A. Borzì and V. Schulz, *Computational optimization of systems governed by partial differential equations*, vol. 8 of Computational Science & Engineering, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012.

[29] F. Bouchut, *Smoothing effect for the non-linear Vlasov-Poisson-Fokker-Planck system*, J. Differ. Equ., 122 (1995), pp. 225–238.

[30] G. E. P. Box and M. E. Muller, *A note on the generation of random normal deviates*, Ann. Math. Statist., 29 (1958), pp. 610–611.

[31] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011.

[32] R. W. Brockett, *Minimum attention control*, in Proceedings of the 36th IEEE Conference on Decision and Control, vol. 3, IEEE, 1997, pp. 2628–2632.

[33] ——, *Optimal control of the Liouville equation*, in Proceedings of the International Conference on Complex Geometry and Related Fields, vol. 39 of AMS/IP Stud. Adv. Math., Amer. Math. Soc., Providence, RI, 2007, pp. 23–35.

[34] ——, *Notes on the control of the Liouville equation*, in Control of partial differential equations, vol. 2048 of Lecture Notes in Math., Springer, Heidelberg, 2012, pp. 101–129.

[35] A. Burshtein, *Introduction to Thermodynamics and Kinetic Theory of Matter*, A Wiley-Interscience publication, Wiley, 1996.

[36] R. Caflisch, D. Silantyev, and Y. Yang, *Adjoint dsmc for nonlinear boltzmann equation constrained optimization*, J. Comput. Phys., (2021), p. 110404.

[37] R. E. Caflisch, *Monte Carlo and quasi-Monte Carlo methods*, vol. 7 of Acta Numer., Cambridge Univ. Press, Cambridge, 1998.

[38] E. J. Candès, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.

[39] G. Cavagnari, A. Marigonda, and B. Piccoli, *Averaged time-optimal control problem in the space of positive Borel measures*, ESAIM Control Optim. Calc. Var., 24 (2018), pp. 721–740.

[40] C. Cercignani, *On boltzmann equation with cutoff potentials*, Phys. Fluids, 10 (1967), pp. 2097–2104.

[41] ——, *Mathematical methods in kinetic theory*, Plenum Press, New York, 1969.

[42] ——, *The Boltzmann Equation and Its Applications*, vol. 67 of Applied Mathematical Sciences, Springer-Verlag, New York, 1988.

[43] C. Cercignani, R. Illner, and M. Pulvirenti, *The Mathematical Theory of Dilute Gases*, vol. 106 of Applied Mathematical Sciences, Springer-Verlag, New York, 1994.

[44] S. Chapman, T. G. Cowling, and D. Burnett, *The mathematical theory of non-uniform gases: An account of the kinetic theory of viscosity, thermal conduction, and diffusion in gases*, Cambridge University Press, New York, 1990.

[45] A. Chertock and A. Kurganov, *On splitting-based numerical methods for convection-diffusion equations*, in Numerical methods for balance laws, vol. 24 of Quad. Mat., Dept. Math., Seconda Univ. Napoli, Caserta, 2009, pp. 303–343.

[46] A. Chertock, A. Kurganov, and G. Petrova, *Fast explicit operator splitting method for convection-diffusion equations*, Internat. J. Numer. Methods Fluids, 59 (2009), pp. 309–332.

[47] H. Cho, D. Venturi, and G. E. Karniadakis, *Numerical methods for high-dimensional probability density function equations*, J. Comput. Phys., 305 (2016), pp. 817–837.

[48] G. Ciaramella and A. Borzì, *A LONE code for the sparse control of quantum systems*, Comput. Phys. Commun., 200 (2016), pp. 312–323.

[49] G. Ciaramella and A. Borzì, *Quantum optimal control problems with a sparsity cost functional*, Numer. Funct. Anal. Optim., 37 (2016), pp. 938–965.

[50] D. Cinquegrana, R. Votta, C. Purpura, and E. Trifoni, *Continuum breakdown and surface catalysis effects in nasa arc jet testing at scirocco*, Aerosp. Sci. Technol., 88 (2019), pp. 258–272.

[51] F. Clarke, *Functional analysis, calculus of variations and optimal control*, vol. 264 of Graduate Texts in Mathematics, Springer, London, 2013.

[52] F. H. Clarke, *Optimization and nonsmooth analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley & Sons, Inc., New York, 1983.

[53] R. Clausius, *X. on the mean length of the paths described by the separate molecules of gaseous bodies on the occurrence of molecular motion: together with some other remarks upon the mechanical theory of heat*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 17 (1859), pp. 81–91.

[54] P. Cockshott and D. Zachariah, *Conservation laws, financial entropy and the eurozone crisis*, Economics, 8 (2014).

[55] M. Coco, A. Majorana, G. Nastasi, and V. Romano, *High-field mobility in graphene on substrate with a proper inclusion of the Pauli exclusion principle*, Atti Accad. Peloritana Pericolanti Cl. Sci. Fis. Mat. Natur., 97 (2019), pp. A6, 15.

[56] M. Coco, G. Mascali, and V. Romano, *Monte Carlo analysis of thermal effects in monolayer graphene*, J. Comput. Theor. Transp., 45 (2016), pp. 540–553.

[57] G. Colonna and A. D'Angola, *Plasma Modeling: Methods and Applications*, IOP Publishing, New York, 2016.

[58] S. D. Conte and C. de Boor, *Elementary Numerical Analysis*, McGraw-Hill Book Co., 1980.

[59] G. Crippa, *The flow associated to weakly differentiable vector fields*, vol. 12 of Tesi. Scuola Normale Superiore di Pisa (Nuova Series) [Theses of Scuola Normale Superiore di Pisa (New Series)], Edizioni della Normale, Pisa, 2009.

[60] J. Delgado and M. Ruzhansky, *Schatten classes on compact manifolds: kernel conditions*, J. Funct. Anal., 267 (2014), pp. 772–798.

[61] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[62] ——, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[63] R. Duan, *Stability of the Boltzmann equation with potential forces on torus*, Phys. D, 238 (2009), pp. 1808–1820.

[64] J. J. Duderstadt and W. R. Martin, *Transport Theory*, John Wiley & Sons, New York-Chichester-Brisbane, 1979. A Wiley-Interscience Publication.

[65] S. Duhr and D. Braun, *Why molecules move along a temperature gradient*, Proc. Natl. Acad. Sci. U.S.A., 103 (2006), pp. 19678–19682.

[66] L. C. Evans, *Partial Differential Equations*, vol. 19 of Graduate Studies in Mathematics, Amer. Math. Soc., Providence, RI, second ed., 2010.

[67] E. Feireisl and A. Novotný, *Singular limits in thermodynamics of viscous fluids*, Advances in Mathematical Fluid Mechanics, Birkhäuser Verlag, Basel, 2009.

[68] K. A. Fichthorn and W. H. Weinberg, *Theoretical foundations of dynamical Monte Carlo simulations*, J. Chem. Phys, 95 (1991), pp. 1090–1096.

[69] B. Fornberg, *Generation of finite difference formulas on arbitrarily spaced grids*, Math. Comp., 51 (1988), pp. 699–706.

[70] A. L. Garcia, *Numerical Methods for Physics*, Prentice Hall Englewood Cliffs, New Jersey, 2000.

[71] D. K. Gathungu and A. Borzì, *A multigrid scheme for solving convection-diffusion-integral optimal control problems*, Comput. Vis. Sci., 22 (2019), pp. 43–55.

[72] M. F. Gelin, A. P. Blokhin, V. A. Tolkachev, and W. Domcke, *Microscopic derivation of the Keilson – Storer master equation*, J. Chem. Phys., 462 (2015), pp. 35 – 40.

[73] M. F. Gelin and D. S. Kosov, *Molecular reorientation in hydrogen-bonding liquids: Through algebraic t - 3/2 relaxation toward exponential decay*, J. Chem. Phys., 124 (2006), p. 144514.

[74] T. Gerya, *Numerical solutions of the momentum and continuity equations*, in Introduction to Numerical Geodynamic Modelling, Cambridge University Press, 2019, pp. 82–104.

[75] A. Glikson, *Theory of existence and uniqueness for the nonlinear Maxwell-Boltzmann equation. I*, Bull. Austral. Math. Soc., 16 (1977), pp. 379–414.

[76] E. Godlewski and P.-A. Raviart, *Hyperbolic systems of conservation laws*, vol. 3/4 of Mathématiques & Applications (Paris) [Mathematics and Applications], Ellipses, Paris, 1991.

[77] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*, Mathematics of computation, 67 (1998), pp. 73–85.

[78] H. Grad, *Asymptotic theory of the boltzmann equation*, Phys. Fluids, 6 (1963), pp. 147–181.

[79] I. Grattan-Guinness, ed., *Landmark writings in western mathematics 1640–1940*, Elsevier B. V., Amsterdam, 2005.

[80] Q. Guo, M. K. Mandal, and M. Y. Li, *Efficient hodge–helmholtz decomposition of motion fields*, Pattern Recognit. Lett., 26 (2005), pp. 493–501.

[81] J. Haack, C. Hauck, C. Klingenberg, M. Pirner, and S. Warnecke, *A consistent bgk model with velocity-dependent collision frequency for gas mixtures*, arXiv preprint arXiv:2101.09047, (2021).

[82] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration illustrated by the Störmer-Verlet method*, Acta Numer., 12 (2003), pp. 399–450.

[83] G. Herdrich, C. Syring, T. Torgau, Y. Chan, and D. Petkow, *An approach for thrust and losses in inertial electrostatic confinement devices for electric propulsion applications*, in 34th International Electric Propulsion Conference, 2015.

[84] G. Herdrick and D. Petkow, *High-enthalpy, water-cooled and thin-walled ICP sources characterization and MHD optimization*, J. Plasma Phys., 74 (2008), p. 391–429.

[85] C. Jacoboni, *Theory of Electron Transport in Semiconductors: A Pathway from Elementary Physics to Nonequilibrium Green Functions*, vol. 165, Springer Science & Business Media, 2010.

[86] J. Keilson and J. E. Storer, *On Brownian motion, Boltzmann's equation, and the Fokker-Planck equation*, Quart. Appl. Math., 10 (1952), pp. 243–253.

[87] C. Kim and D. Lee, *The Boltzmann equation with specular boundary condition in convex domains*, Comm. Pure Appl. Math., 71 (2018), pp. 411–504.

[88] D. S. Kosov, *Telegraph noise in Markovian master equation for electron transport through molecular junctions*, J. Chem. Phys., 148 (2018).

[89] M. Kraposhin, A. Bovtrikova, and S. Strijhak, *Adaptation of kurganov-tadmor numerical scheme for applying in combination with the PISO method in numerical simulation of flows in a wide range of mach numbers*, Procedia Computer Science, 66 (2015), pp. 43–52.

[90] F. Y. Kuo and I. H. Sloan, *Lifting the curse of dimensionality*, Notices Amer. Math. Soc., 52 (2005), pp. 1320–1329.

[91] A. Kurganov and G. Petrova, *A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system*, Commun. Math. Sci., 5 (2007), pp. 133–160.

[92] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.

[93] K. Lakshmi, R. Parvathy, S. Soumya, and K. Soman, *Image denoising solutions using heat diffusion equation*, in 2012 International Conference on Power, Signals, Controls and Computation, IEEE, 2012, pp. 1–5.

[94] J.-L. Lions, *Optimal control of systems governed by partial differential equations*, Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170, Springer-Verlag, New York-Berlin, 1971.

[95] A. J. Lofthouse, I. D. Boyd, and M. J. Wright, *Effects of continuum breakdown on hypersonic aerothermodynamics*, Phys. Fluids, 19 (2007), p. 027105.

[96] B. Luan and M. O. Robbins, *The breakdown of continuum models for mechanical contacts*, Nature, 435 (2005), pp. 929–932.

[97] R. Mazo, *Brownian Motion: Fluctuations, Dynamics, and Applications*, Oxford University Press, 2002.

[98] A. Montanaro, *Quantum speedup of Monte Carlo methods*, Proc. A., 471 (2015), pp. 20150301, 20.

[99] D. Montgomery, *Brownian motion from Boltzmann's equation*, Phys. Fluids, 14 (1971), pp. 2088–2090.

[100] E. P. Muntz, *Rarefied gas dynamics*, in Annual review of fluid mechanics, Vol. 21, Annual Reviews, Palo Alto, CA, 1989, pp. 387–417.

[101] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics*, The Clarendon Press, Oxford University Press, 1999.

[102] H. Nishikawa, *A truncation error analysis of third-order muscl scheme for nonlinear conservation laws*, International Journal for Numerical Methods in Fluids, (2020).

[103] S. Osher, *Convergence of generalized MUSCL schemes*, SIAM J. Numer. Anal., 22 (1985), pp. 947–961.

[104] S. Osher and S. Chakravarthy, *High resolution schemes and the entropy condition*, SIAM J. Numer. Anal., 21 (1984), pp. 955–984.

[105] W. F. Phillips, *Motion of aerosol particles in a temperature gradient*, Phys. Fluids, 18 (1975), pp. 144–147.

[106] N. Pogodaev, *Optimal control of continuity equations*, Nonlinear Differ. Equ. Appl., 23 (2016), pp. Art. 21, 24.

[107] M. Pulvirenti, C. Saffirio, and S. Simonella, *On the validity of the Boltzmann equation for short range potentials*, Rev. Math. Phys., 26 (2014), pp. 1450001, 64.

[108] L. Q. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[109] H. Risken, *The Fokker-Planck equation*, vol. 18 of Springer Series in Synergetics, Springer-Verlag, Berlin, second ed., 1989. Methods of solution and applications.

[110] S. Rjasanow and W. Wagner, *Stochastic Numerics for the Boltzmann Equation*, vol. 37 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2005.

[111] V. Romano, A. Majorana, and M. Coco, *DSMC method consistent with the Pauli exclusion principle and comparison with deterministic solutions for charge transport in graphene*, J. Comput. Phys., 302 (2015), pp. 267–284.

[112] W. Rudin, *Real and Complex Analysis*, McGraw-Hill Book Co., New York, third ed., 1987.

[113] V. Rusanov, *Calculation of intersection of non-steady shock waves with obstacles*, J. Comput. Math. Phys. USSR, 1 (1961), pp. 267–279.

[114] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[115] R. D. Skeel, G. Zhang, and T. Schlick, *A family of symplectic integrators: stability, accuracy, and molecular dynamics applications*, SIAM J. Sci. Comput., 18 (1997), pp. 203–222.

[116] I. H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.

[117] R. F. Snider, *Eigenvectors and eigenvalues for the Keilson-Storer collision kernel*, Phys. Rev. A, 33 (1986), pp. 178–181.

[118] R. L. Speth, W. H. Green, S. MacNamara, and G. Strang, *Balanced splitting and rebalanced splitting*, SIAM J. Numer. Anal., 51 (2013), pp. 3084–3105.

[119] G. Stadler, *Elliptic optimal control problems with L^1-control cost and applications for the placement of control devices*, Comput. Optim. Appl., 44 (2009), pp. 159–181.

[120] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[121] M. L. Strekalov, *Population relaxation of highly rotationally excited molecules at collisions*, Chem. Phys. Lett., 548 (2012), pp. 7 – 11.

[122] H. Struchtrup, *Macroscopic transport equations for rarefied gas flows*, Interaction of Mechanics and Mathematics, Springer, Berlin, 2005. Approximation methods in kinetic theory.

[123] T. Tang and Z. Teng, *Monotone difference schemes for two dimensional nonhomogeneous conservation laws*, Pitman Research Notes in Mathematics Series, (1998), pp. 229–243.

[124] L. Tartar, *An introduction to Sobolev spaces and interpolation spaces*, vol. 3 of Lecture Notes of the Unione Matematica Italiana, Springer, Berlin; UMI, Bologna, 2007.

[125] H. Tran, J.-M. Hartmann, F. Chaussard, and M. Gupta, *An isolated line-shape model based on the Keilson-Storer function for velocity changes. ii. molecular dynamics simulations and the q(1) lines for pure h2*, J. Chem. Phys., 131 (2009), p. 154303.

[126] F. Tröltzsch, *Optimal control of partial differential equations. Theory, methods and applications*, vol. 112 of Graduate Studies in Mathematics, Amer. Math. Soc., Providence, RI, 2010.

[127] M. Ulbrich, *Semismooth Newton methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, vol. 11 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[128] M. Urabe, *Theory of errors in numerical integration of ordinary differential equations*, Journal of Science of the Hiroshima University, Series A-I (Mathematics), 25 (1961), pp. 3 – 62.

[129] N. G. van Kampen, *Stochastic processes in physics and chemistry*, vol. 888 of Lecture Notes in Mathematics, North-Holland Publishing Co., Amsterdam-New York, 1981.

[130] L. Verlet, *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules*, Phys. Rev., 159 (1967), p. 98.

[131] W. Wagner, *A convergence proof for Bird's direct simulation Monte Carlo method for the Boltzmann equation*, J. Statist. Phys., 66 (1992), pp. 1011–1044.

[132] M. D. Weinberg, *Direct simulation monte carlo for astrophysical flows–i. motivation and methodology*, Mon. Notices Royal Astron. Soc., 438 (2014), pp. 2995–3006.

[133] J. Welty, G. L. Rorrer, and D. G. Foster, *Fundamentals of momentum, heat, and mass transfer*, John Wiley & Sons, New Jersey, 2020.

[134] R. D. White, R. Robson, S. Dujko, P. Nicoletopoulos, and B. Li, *Recent advances in the application of boltzmann equation and fluid equation methods to charged particle transport in non-equilibrium plasmas*, J. Journal of Physics D, 42 (2009), p. 194001.