Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
MATEMATICA E INFORMATICA

CICLO XXVI

COORDINATORE Prof. Massimiliano Mella

KINETIC APPROXIMATION, STABILITY AND CONTROL OF COLLECTIVE BEHAVIOR IN SELF-ORGANIZED SYSTEMS

Settore Scientifico Disciplinare MAT/08

Dottorando	**Tutore**
Dott. ALBI GIACOMO	Prof. PARESCHI LORENZO

Anni 2011/2013

To my family

Contents

Preface

> *"Look deep into nature, and then you will understand everything better."*
>
> Albert Einstein

Rome, outside the main station, a huge flock of birds flies around, shaping in ever-changing geometrical figures. Philippine, some meters below the sea level, a shoal of fishes creates a massive rotating body, scaring any potential predator. People can only stare whit amazed eyes at these natural phenomena and at their incredible level of organization, as if a superior intelligence steers them. The beauty of these phenomena may justify a detailed study, but the main motivation arises from observing that the property of a group of agents to create ordered patterns is non-trivial. Aristoteles outlined this concept in his famous quote "*the whole is greater then the sums of the parts*", and this idea has evolved across ages till the modern definition of *self-organized system*, namely a system that exhibits a spontaneous order out of a multiplicity of simple interactions. Its first formalization was proposed by the cybernetician W. R. Ashby in [13, 14] and is deeply interconnected with the

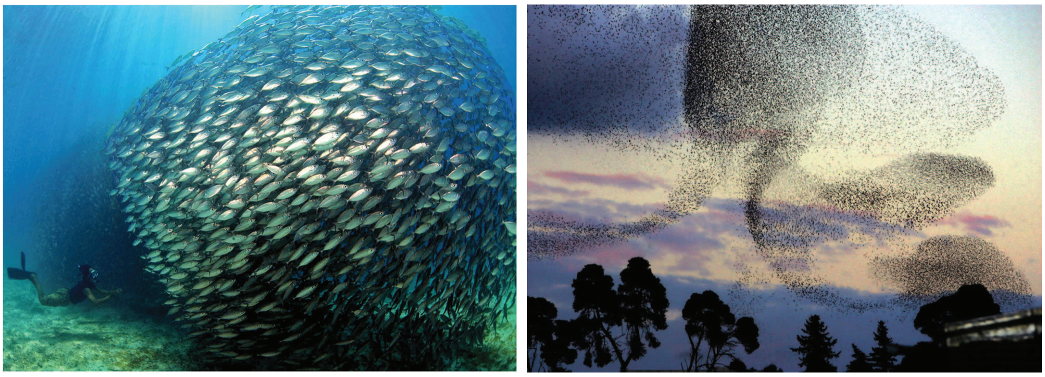


Figure 1: On the left, a school of fishes shows a typical milling behavior on the right, murmuration in Rome.

idea of *emergence behavior*, as "*the arising of novel and coherent structures, patterns and properties during the process of self-organization [. . .]*", [63].

Many natural systems show emerging properties, such as cells, swarming dynamics, galaxy formations, chemical compounds, organisms or crystals; but we can also identify self-organized systems in human artifacts, for example in market economies, crowd dynamics, vehicular traffics, opinion formations, wealth distributions, networks, cybernetics or artificial intelligences.

The scientific community has been therefore attracted by the study of self-organized systems, in various areas such as biology [105, 45], physics [158, 103], mathematics [47, 75, 66, 23, 24], engineer [91, 15], computer science [35, 150] , economics and sociology [130, 154], thus designing a fruitful and multidisciplinary research field.

Figure 2: On the left, computer animation of a swarm from the movie *The Croods*, on the right fleet of robots from the open source project swarmrobot.org.

This research field has been exponentially growing in the last decades, since massive computing techniques allowed more powerful analysis of self-organized systems and for the development of several applications. At the end of 80s, *Boids* simulations by C. Reynolds[1], [150], shaded light on the usage of a simple swarming model, then extensively developed in computer graphics for movies and video games, [76, 163]. New perspectives arose with the concept of *swarming intelligence*, [35], the main idea was to use the self-organization property of swarming system to perform complex tasks, for example in optimization, [116], or in engineer with swarm robotics, where an ensemble of small size robots performs particular actions, improving efficiency and cost saving, [26, 126]. Other, civil engineers are interested in crowds models to test the load of structures and to plan evacuation strategy in buildings [69, 165]. In aerospace engineering, synchronization systems for flight formations of several spacecraft missions (DARWIN, CloudSat, CanX), based on flocking models [40, 148, 136].

[1] http://www.red3d.com/cwr/boids/

The large interest in self-organized system led also to the development of novel techniques, models and theorems in mathematics. From a mathematical point of view, a description of self-organized models is provided by complex system theory, where the overall dynamic is depicted by a nonlinear ODEs system. In the following we will refer to these models as *microscopic description level*.

Microscopic models describing the evolution of a population of agents are usually called individual based models or interacting multi-agent systems. The different levels of dimensionality and complexity these models present are related to the behavior under investigation.

In terms of socio-economic dynamics, such as consensus formation or wealth exchange, we study the evolution of a single property, e.g. the agent income or opinion, whereas in terms of animal, bacteria or crowd behaviors, we most likely look to the variation of velocity and position of the agents, which usually corresponds to a second order ODEs system. The aim of this thesis will mainly focus on these two types of dynamics, but related models with higher dimensionality and order can be included, as modeling the behavior of market customers or social-network users potentially requires the evolution of more than two single properties.

In general microscopic models for interacting agents have the same structure of many classical problems in computational physics which require the evaluation of all pairwise long range interactions in a large ensembles of particles. The N-body problem of gravitation (or electrostatics) is a classical example. Such problem involves the evaluation of summations of the type

$$S_i^N = \sum_{j=1}^{N} w_j K(x_i, x_j), \qquad \forall\ i.$$

A direct evaluation of such sums at N target points clearly involve a $O(N^2)$ cost, therefore study of microscopic model for a large system of individuals implies a considerable effort in numerical simulations, as microscopic models based on real data may take into account very large numbers of interacting individuals, [67, 158].

A step towards the reduction of computational complexity of the microscopic model is represented by the idea of the application of more general level of descriptions, which has been extensively developed in the kinetic research research (see [138, 75, 53, 72, 73, 100]) and it implies that the derivation of *mesoscopic* and *macroscopic models* presents a first approximation of the original dynamic.

The basic idea of *kinetic equation* lies on a different level of phenomena description: instead of focusing on the evolution of single particles, it analyzes the density of particles as in the classical Boltzmann gas dynamic. The interest of kinetic equation arises in several disciplines, in astrophysics for galactic dynamics, in molecular biology for chemotaxis, in plasma physics [77], medical physics for radiotherapy [110] and animal swarming [53].

A rigorous derivation of a kinetic model from the microscopic particle system constitutes a mathematical issue, also in the case of the Boltzmann equation the rigorous limit holds just for a short time, but not global result it is known yet, [122]. For interacting multiagent systems in swarming and flocking different approaches have been used, like BBGKY hierarchy [100], or *mean field limit* [47], or the *binary interaction* approximation, [155, 149]. Certainly by passing from a microscopic description based on phase-space particles $(x_i(t), v_i(t))$ to a mesoscopic level where the object of study is a particle distribution function $f(x, v, t)$ redefines the model in a new one.
Specifically kinetic equations for interacting agent system are described typically by Vlasov-type equations or Fokker–Planck equations, in presence of noise effects. Such context usually involve high dimensional and nonlinear terms opening several directions on the numerical approach to use, [86, 98, 6].

The aim of this thesis is to design new perspectives in kinetic modeling of self-organized systems, with particular attention to the development of numerical methods and extensions to control dynamics. Each chapter is self consistent, with its own introduction, results and conclusion, each one referring to a research article, already published or under revision process in peer-review journals.

Chapter 1 reports the research made in [7], jointly with Lorenzo Pareschi, where inspired by the techniques introduced in [34] for plasma physics, we develop a direct simulation Monte Carlo methods based on a binary collision dynamic described by the corresponding kinetic equation. The theoretical ground of the algorithms is represented by the stochastic approximation through the Boltzmann-Povzner formula of the mean-field dynamic for a general swarming models, [53], which reads

$$\partial_t f + v \cdot \nabla_x f = -\nabla_v \cdot (\xi[f]f).$$

The method developed permits to approximate the microscopic dynamic of N particles at a cost directly proportional to the number of sample particles, N_s, involved in the computation, thus avoiding the quadratic computational cost. Furthermore, in contrast with classical methods for Boltzmann equation [34], the nature of the approximating equations is such that the resulting Monte Carlo algorithms are fully *meshless*, due to the long-range interactions of classical swarming models.
In order to make a step further in the direction of more realistic swarming models, several features have been taken into account, for example *roosting forces*, *topological interactions*, *limited perception* and more refined dynamics [51, 19, 138, 53]. Several numerical results show the efficiency of the algorithms proposed.

Classical multiagent interacting models for sociological or biological phenomena are based on the so called *three zone model*, which enlightens three main interaction rules: *attraction*, *repulsion* and *consensus*. Just by these simple interaction dynamics complex behaviors emerge and from the analytical point of view interesting questions arise in terms of model solution stability [28, 80].

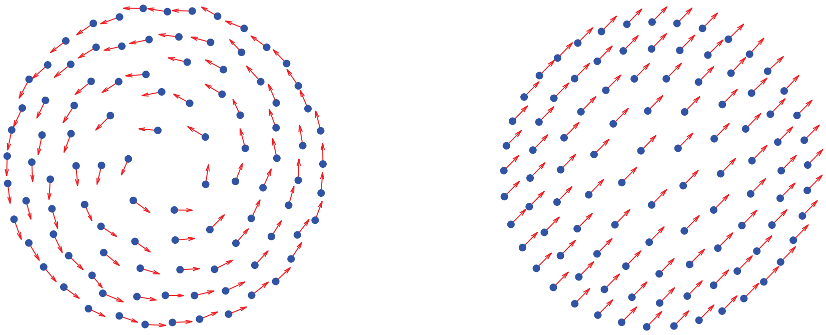


Figure 3: Milling and flocking patterns arising in swarming models with attraction/repulsion dynamics.

Chapter 2 reports the joint work [2], with José A. Carrillo, Daniel Balagué, James Von Brecht, where we address our interest to the linear stability of particular solutions of second order individual based models for biological swarming, called *flock ring* and *mill ring* solutions.
The individuals interact via a nonlocal interaction potential that is repulsive in the short range and attractive in the long range, which in general reads

$$\begin{aligned}\dot{x}_i &= v_i\\ \dot{v}_i &= S(x_i, v_i) + \frac{1}{N}\sum_{j\neq i}\nabla W(x_j - x_i),\end{aligned}$$

for particular choices of function $S(\cdot,\cdot)$ and we relate the instability of the flock rings with the instability of the ring solution of the first order model, of the form

$$\dot{x}_i = \frac{1}{N}\sum_{j\neq i}\nabla W(x_j - x_i).$$

We observe that repulsive-attractive interactions lead to clustering and fattening instabilities for flock rings that prove analogous to similar instabilities that occur for ring solutions of the first order model. Finally, we numerically explore mill patterns arising from these interactions by varying the asymptotic speed of the system. The results of this chapter have been extended in [55], for the case of non linear stability of *flock ring* solutions.

Chapter 3 develops a framework for the description of a group of large number of agents, influenced by a small number of external individuals. The result refers to [6], a joint research with Lorenzo Pareschi, where we start from the microscopic dynamic, deriving two other different levels of description: the mesoscopic (or kinetic) level through a mean-field limit and the macroscopic level through a suitable hydrodynamic approximation. In both cases the resulting dynamic appears as coupled system of PDEs and ODEs, similar settings has been studied in [81] for opinion formation and for pedestrian dynamic in [66].

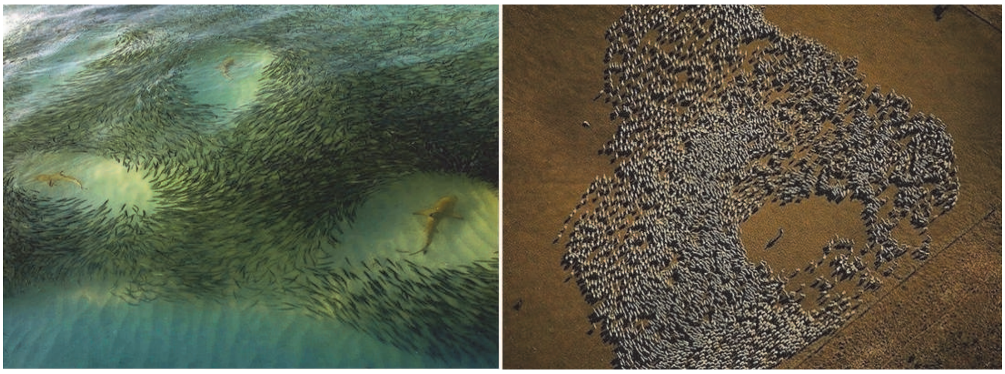


Figure 4: On the left a shoal of fishes reacting to shark attacks, on the right sheepdogs regrouping the herd.

In a biological context, this corresponds to the behavior of a flock of birds or a school of fish attacked by one or more predators, or the movement of a herd of sheep guided by a sheepdog. From the modeling viewpoint this involves a microscopic dynamic described by classical flocking models interacting with a set of few individuals characterized in way similar to what was done in [59]. Moreover, we endow the classical dynamic of interaction both with a metric as well as a topological interaction rule, [19].

From a general view point the idea presented in Chapter 3 can be interpreted as a first effort to control a self-organized system through the presence of an external dynamic. A natural improvement of this approach implies the use of optimal control theory to steer the system to purse a desired state. Such problems have been studied initially in engineer and computer science communities, in particular for applications in robotics [91, 128, 137]. Most recently mathematics have focused their attention to these problems from different point of views: at the level of microscopic [37], for mean-field models [48, 88, 87, 27] and in conservation laws models in [60].

In a joint research with M. Herty and L. Pareschi, in [5], we worked on feedback control of such processes, which can be used to study the exterior influence of the system dynamics. We report the results of this approach in Chapter 4, where an optimal control problem for a large system of interacting agents is considered using a kinetic perspective. As a prototype model we analyze a microscopic model of opinion formation under constraints, which reads as follows

$$\dot{w}_i = \frac{1}{N}\sum_{j=1}^{N} P(w_j, w_i)(w_j - w_i) + u,$$
$$u = \operatorname{argmin} \int_0^T \left(\frac{1}{N}\sum_{j=1}^{N}(w_j - w_d)^2 + \frac{\nu}{2}u^2 \right) dt,$$

where $w \in \mathcal{I} = [-1, 1]$, for some choice of $P(\cdot,\cdot)$ and w_d. The aim of the chapter is to give a kinetic description of this optimal control problem, therefore a Boltzmann-type equation based on a *model predictive control* formulation is introduced and discussed. The main difference respect to classical swarming models is introduced by the bounds of w, which must be preserved by, the Boltzmann equation. In particular, the receding horizon strategy permits to embed the minimization of suitable cost functional into binary particle interactions, where a noise component is added. The corresponding Fokker-Planck asymptotic limit is derived and reads and explicit expressions of stationary solutions are given.

The last part of Chapter 4 extends the methodology used to the non-homogeneous case, whereas optimal control problem is solved for mean-field flocking models. Optimal control problems for refined flocking models have been recently investigated in a similar setting at kinetic level in, [88, 49], introducing the concept of sparse control and for microscopic model [37], where the idea is to control the swarm through the action of a population of leaders. Here we derive a corresponding kinetic approximation of the control problem, through the mean-field limit. Extensions to a kinetic description of optimal control problem through leaders are under investigation.

In Chapter 5 we are also interested in optimal control problems for kinetic equations and report a recent result obtained [3], together with M. Herty, C. Jörres, L. Pareschi. Many applications, from aerospace and mechanical engineering to the life sciences, involve a systems of differential equations of the form

$$y'(t) = f(y(t), t) + \frac{1}{\varepsilon} g(y(t), t),$$

in particular the development of numerical methods for time discretization of optimal control problems involving differential equations has been an intensive field of research [101, 107].

More precisely, we consider the development of implicit-explicit (IMEX) time integration schemes for optimal control problems of boundary problems governed by the *Goldstein–Taylor* model. In the diffusive scaling this model is a hyperbolic approximation to the heat equation. We investigated the relation of time integration schemes and the formal Chapman-Enskog type limiting procedure. For the class of stiffly accurate implicit-explicit Runge-Kutta methods (IMEX) the discrete optimality system also provides a stable numerical method for optimal control problems governed by the heat equation. The methodology presented opens new perspectives to extend the same techniques for radiative transfer equation, which gives a description of radiotherapy process in biomedical applications [110, 4].

CHAPTER 1

Binary interaction algorithms for flocking and swarming dynamics

1.1 Introduction

The study of mathematical models describing collective behavior and synchronized motion of animals, like bird flocks, fish schools and insect swarms, has attracted a lot of attention in recent years [1, 50, 66, 54, 141, 67, 80, 138, 158, 47]. In biological systems such behaviors are observed in every level of the food chain, from the swarm intelligence of the zoo plankton, to bird flocking and fish schools, to mammals moving in formation [84, 113, 142]. Beside biology, emerging collective behaviors play a relevant role in several applications involving the dynamics of a large number of individuals/particles which range from computer science [150], physics [98] and engineering [124] to social sciences and economy [46]. We refer to [141] for a recent review of some of the mathematical topics and the applications involved.

Naturally occurring synchronized motion has inspired several directions of research within the control community. A well-known application is related to formation flying missions and missions involving the coordinated control of several autonomous vehicles [120]. There are several current projects which are dealing with the formation flying and coordinated control of satellites, like the DARWIN project of the European Space Agency (ESA) with the goal of launching a space-based telescope aiding in the search for possible life-supporting planets, or the PRISMA project led by the Swedish Space Corporation (SSC) which will be the first real formation flying space mission launched [70].

In this manuscript we will focus on general models which are capable to reproduce flocking, swarming and other collective behaviors. Most of the classical models describing these phenomena are based on the simple definition of three interacting zones, the so called *three-zones model* [10, 114].

Let us briefly summarize the three-zones assumptions. We define three regions around each individual: a short-range repulsion zone, an intermediate velocity alignment zone and a long-range attraction zone (see Figure 1.1). Each interaction between individuals is evaluated accordingly to the relative position in the model.

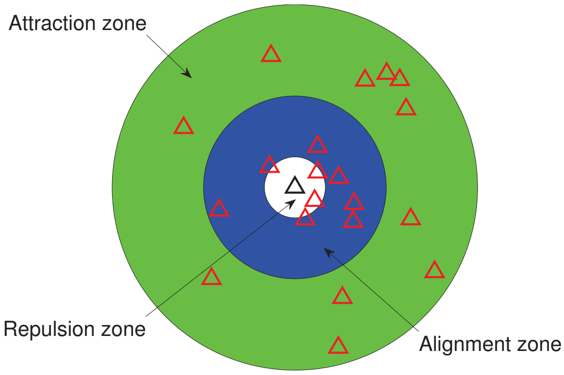


Figure 1.1: Sketch of the three-zone model.

- *Repulsion zone:* when individuals are too close each other they tend to move away from that area.

- *Alignment zone:* individuals try to identify the possible direction of the group and to align with it.

- *Attraction zone:* when individuals are too far from the group they want to get closer.

Typically different interaction models are taken in the different zones [1, 80] or the modelling is focused on a specific zone, like the alignment/consensus dynamic [67, 138]. Of course the particular shape and size of the zones depends on the specific application considered. For example recent studied on birds flocks suggest that each bird modifies its position, relative to few individuals directly surrounding it, no matter how close or how far away those individuals are [19]. It is not clear however if this applies also to other kind of animals.

Studying this kind of dynamics for large system of individuals implies a considerable effort in numerical simulations, microscopic models based on real data my take into account very large numbers of interacting individuals (from several hundred thousands up to millions). Computationally the problems have the same structure of many classical problems in computational physics which require the evaluation of all pairwise long range interactions in a large ensembles of particles.

The N-body problem of gravitation (or electrostatics) is a classical example. Such problems involve the evaluation of summations of the type

$$S_i^N = \sum_{j=1}^{N} w_j K(x_i, x_j), \qquad \forall\ i. \tag{1.1}$$

A direct evaluation of such sums at N target points clearly requires $O(N^2)$ operations and algorithms which reduce the cost to $O(N^\alpha)$ with $1 \leqslant \alpha < 2$, $O(N \log N)$ etc. are referred to as *fast summation methods.* For uniform grid data the most famous of these is certainly the Fast Fourier Transform (FFT). In the case of general data most fast summation methods are approximate methods based on analytical considerations, like the *Fast Multipole* method [98], *Wavelets Transform* methods [30] and, more recently, dimension reduction using *Compressive Sampling* techniques [46], or based on some *Monte Carlo* strategies at different levels [43]. Extensions of the above mentioned approaches to kinetic equations are discussed for example in [140, 125, 9, 86, 143].

From a mathematical modeling point of view, these problem have been developed extensively in the kinetic research community (see [72, 100, 54]) where the derivation of kinetic and an hydrodynamic equations represent a first step towards the reduction of the computational complexity. Of course, passing from a microscopic description based on phase-space particles $(x_i(t), v_i(t))$ to a mesoscopic level where the object of study is a particle distribution function $f(x, v, t)$ redefines the model in a new one where new methods of solution are required.

In this paper we are going to follow this research path in two main directions: first we review the derivation of the different kinetic approximations from the original microscopic models and then we introduce and analyze several stochastic Monte Carlo methods to approximate the kinetic equations. Monte Carlo methods are the most well-known approach for the numerical solution of the Boltzmann equation of rarefied gases in the short-range interactions, and many efficient algorithms have been presented [32, 16, 43, 143]. On the other hand the literature on efficient Monte Carlo strategies for long-range interactions, and thus Landau-Fokker-Plank equations, is much less developed but of great interest in the field of plasma physics [34, 77].

Here, inspired by the techniques introduced in [34, 77] for plasma, we develop direct simulation Monte Carlo methods based on a binary collision dynamic described by the corresponding kinetic equation. The methods permit to approximate the microscopic dynamic at a cost directly proportional to the number of sample particles involved in the computation, thus avoiding the quadratic computational cost. The limiting behavior characterizing the mean-field interaction process of the particles system is recovered under a suitable asymptotic scaling of the binary collision process. In such a limit we show that the Monte Carlo methods here developed are in very good agreement with the direct evaluation of the original microscopic model but with a considerable gain of computational efficiency.

The rest of the manuscript is organized as follows. In the first section we present some of the classical microscopic models for flocking and swarming. Generalization of the notion of visual cone [54] are also discussed. Since the interaction is non local, the derivation of the limiting mean-field kinetic equation is made through a *Povzner-Boltzmann* kinetic equation [149] in the anologous situation of the so-called *grazing collision limit* [53]. To solve the resulting Boltzmann-like mesoscopic partial differential equation we introduce different stochastic binary interaction algorithms and compare their computational efficiency and accuracy with a direct evaluation of the microscopic models and a stochastic approximation of the mean-field kinetic model. We show that the new approach permits to reduce the overall cost from $O(N^2)$ to $O(N)$ operations. In particular we show that the choice $\varepsilon = \Delta t$, where ε is the small scaling parameter leading to the mean field kinetic model, originates binary interaction algorithms consistent with the limiting behavior of the particle system. Furthermore, in contrast with classical methods [34, 77], the nature of the approximating equations is such that the resulting Monte Carlo algorithms are fully mesh less. In the last section of the paper we report several simulations in two and three space dimensions of different microscopic models solved by the binary Monte Carlo method in the above scaling.

1.2 Microscopic models

In this section we review some well-known microscopic models of flocking and swarming (see [67, 80, 138] and the references therein). We are interested in the study of a dynamical system composed of N individuals with the following general structure

$$\begin{cases} \dot{x}_i = v_i, \qquad\qquad i = 1, \dots, N, \\ \\ \dot{v}_i = S(v_i) + \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} \left[H(x_i, x_j)(v_j - v_i) + A(x_i, x_j) + R(x_i, x_j)\right] \psi_\alpha(x_i, x_j, v_i) \end{cases} \tag{1.2}$$

where (x_i, v_i) lives in $\mathbb{R}^{2d}$, $d \geqslant 1$, $S(v_i)$ describes a self-propelling term, $H(x_i, x_j)$ the alignment process, $A(x_i, x_j)$ the attraction dynamic and the term $R(x_i, x_j)$ the short-range repulsion. In (2.1) the multiplicative factor $\psi_\alpha(x_i, x_j, v_i) \in [0, 1]$ takes into account the effects of space perception as a function of some vector of parameters α.

1.2.1 Cucker and Smale model

Cucker and Smale model is a pure alignment model, no repulsion or attraction or other effects are taken in account, see [67, 68] and [53]. The classical model reads as follow

$$\begin{cases} \dot{x}_i = v_i \\ \\ \dot{v}_i = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} H(|x_j - x_i|)(v_j - v_i) \end{cases} \qquad i = 1, \ldots, N, \tag{1.3}$$

where $H(|x_j - x_i|)$ is a function that measures the strength of the interaction between individuals i and j , and depends on the mutual distance, under the assumption that closer individuals have more influence than the far distance ones.

A typical choice of function H is the following

$$H(r) = \frac{K}{(\varsigma^2 + r^2)^\gamma}, \tag{1.4}$$

where $K, \varsigma > 0$ are positive parameters and $\gamma \geqslant 0$. Under this assumptions it can be shown that well-posedness holds for the initial value problem of (1.3) and the solution is mass and momentum preserving, with compact support for position and velocity, see for more details [47, 99].

Moreover in [67, 53] it was established that the parameter γ discriminate the behavior of the solution, in the following way

Theorem 1.2.1 *Let* $\Gamma(t) = \frac{1}{2}\sum_{i\neq j} |x_i(t) - x_j(t)|^2$ *and* $\Lambda(t) = \frac{1}{2}|v_i(t) - v_j(t)|^2$. *If* $\gamma \leqslant \frac{1}{2}$ *then*

(i) exist a positive costant B_0 *such that:* $\Gamma(t) \leqslant B_0$ *for all* $t \in \mathbb{R}$.

(ii) $\Lambda(t)$ *converge towards zero as* $t \to \infty$.

(iii) The vector $x_i - x_j$ *tends to a limit vector* $\hat{x}_{ij}$*, for all* $i, j = 1, \ldots, N$.

In other words, the velocity support collapses exponentially to a single point and the flock holds the same disposition. From this theorem we recover the notion of *unconditional flocking* in the regime $\gamma \leqslant \frac{1}{2}$. If $\gamma > \frac{1}{2}$ in general unconditional flocking doesn't follow, but under some conditions on initial data flocking condition is reached, see [54].

Note that standard Cucker-Smale model prescribes perfectly symmetric interactions and takes in account only the alignment dynamic. As a result total momentum is preserved by the dynamics. The introduction of a limited space perception (like a visual cone) breaks symmetry and momentum conservation. This choice corresponds to take a function for the strength of the interaction of the type

$$H_\alpha(x_i, x_j, v_i) = H(|x_i - x_j|)\psi_\alpha(x_i, x_j, v_i), \tag{1.5}$$

where the parameter vector α is related, for example, to the width of the visual cone.

1.2.2 D'Orsogna-Bertozzi et al. model

The microscopic model introduced by D'Orsogna, Bertozzi et al. [80] considers a self-propelling, attraction and repulsion dynamic and reads

$$\begin{cases} \dot{x}_i = v_i \\ \\ \dot{v}_i = (a - b|v_i|^2)v_i - \dfrac{1}{N}\displaystyle\sum_{j \neq 1} \nabla_{x_i} U(|x_j - x_i|) \end{cases} \qquad i = 1, \dots, N, \tag{1.6}$$

where a, b are nonnegative parameters, $U : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a given potential modeling the short-range repulsion and long-range attraction, and N is the number of individuals. Function U gives us the attraction-repulsion dynamic typically described by a *Morse potential*

$$U(r) = -C_A e^{-r/l_A} + C_R e^{-r/l_R}, \tag{1.7}$$

where C_A, C_R, l_A, l_R are positive constants measuring the strengths and the characteristic lengths of the attraction and repulsion. In (1.6) the term $(a - b|v_i|^2)v_i$ characterizes *self-propulsion* and *friction*. Asymptotically this term give us a desired velocity, in fact for large times the velocity of every single particle tends to $\sqrt{a/b}$.

The most interesting case in biological applications occurs when the constants in the Morse potential satisfy the following inequalities $C := C_R/C_A > 1$ and $l := l_R/l_A < 1$, which correspond to the long range attraction and short range repulsion. Moreover the choice of the parameters fixes the evolution of the N particles system towards a particular equilibrium. The following distinction holds: if $Cl^d > 1$ then crystalline patterns are observed and for $Cl^d < 1$ the motion of particles converges to a circular motion of constant speed, where $d \geqslant 2$ is the space dimension. In [80] a further study of the parameters can be found.

1.2.3 Motsch-Tadmor model

In a recent work [138] the authors propose a modification of the classical Cucker-Smale model as follows

$$\begin{cases} \dot{x}_i = v_i \\ \\ \dot{v}_i = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} h(x_i, x_j)(v_j - v_i), \end{cases} \qquad i = 1, \dots, N, \tag{1.8}$$

where h is defined by

$$h(x_i, x_j) = \frac{H(|x_i - x_j|)}{\bar{H}(x_i)}, \qquad \bar{H}(x_i) = \frac{1}{N}\sum_{k=1}^{N} H(|x_i - x_k|).$$

The model differs from the classical one, since the influence between two particles $H(|x_j - x_i|)$ is weighted by the average influence on the single individual i. In this way the function $h(x_i, x_j)$ loses in general any kind of symmetry property of the original Cucker-Smale dynamic.

We emphasize, however, that in our general setting this model is included in the Cucker-Smale alignment dynamic of type (1.5) with a particular choice for the function defining the space perception of the form

$$\psi_\alpha(x_i, x_j, v_i) = \frac{1}{\bar{H}(x_i)}. \tag{1.9}$$

This can be interpreted as a higher perception level of zones where the individuals have a higher concentration and a lower interest in zones where individuals are more scattered.

1.2.4 Perception cone, topological interactions and roosting force

For interacting animals like birds, fishes, insects the visual perception of the single individual plays a fundamental role [65, 83, 84]. In [54] the authors introduce in the dynamic a further rule: the *visual cone*. A visual cone identifies the area in which interaction is possible and blind area where can not be interaction. Mathematically speaking the visual cone depends on an angle, θ, that give us the visual width. Together with position and velocity the visual area can be described as follows

$$\Sigma(x_i, v_i, \theta) = \left\{ y \in \mathbb{R}^d : \frac{(x_i - y) \cdot v_i}{|(x_i - y)|\,|v_i|} \geqslant \cos(\theta/2) \right\}. \tag{1.10}$$

As already discussed the introduction of a visual cone breaks the typical symmetry of the interaction (see Figure 1.2).

The drawback of this choice is that a single individual that has no one in his visual cone, never changes his direction. For real situations this assumption is clearly too strong, since many other stimuli are received by the surrounding. We cannot ignore other perceptions like hearing, smell and visual memory. For example fishes use their visual perception mostly on large/medium distance whereas on medium/short distance they rely on their lateral line. These observations lead naturally to improve the idea of a visual cone by introducing a *perception cone* as follows: we assign two different weights measuring the strength/probability of the interaction. A weight p_1 in the case of strong perception and p_2 in case of weak perception, with $0 \leqslant p_2 \leqslant p_1 \leqslant 1$. Note that taking $p_1 = 1$ and $p_2 = 0$ we have the standard visual cone. For example, in the simulation section we consider a perception cone ψ_α, $\alpha = (\theta, p_1, p_2)$, with the following form

$$\psi_\alpha(x_i, x_j, v_i) = p_2 + (p_1 - p_2)\mathbb{1}_{\Sigma(x_i, v_i, \theta)}(x_j), \tag{1.11}$$

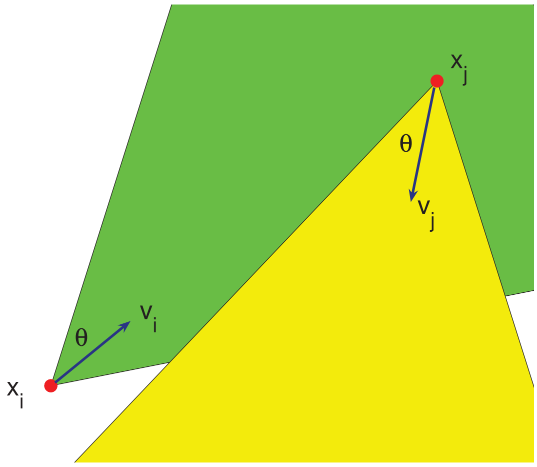


Figure 1.2: One of the possible configurations in the interaction with a perception cone. Individual j is perceived by individual i but not vice versa.

where $\mathbb{1}_{\Sigma(x_i,v_i,\theta)}(\cdot)$ is the indicator function of the set $\Sigma(x_i, v_i, \theta)$ defined according to (1.10).

Related efforts to improve the dynamic consider also different ingredients like *topological interactions* where individuals interact only with the closest individuals and with a limited number of them, see [19, 66]. Another variant concerns the introduction of a term describing a *roosting force* [51, 1]. In fact, flocking phenomena tends to stay localized in a particular area, this force acts orthogonal to the single velocity, giving each particle a tendency towards the origin.

1.3 Kinetic equations

For a realistic numerical simulation of a flock the number of interacting individuals can be rather large, thus we need to solve a very large system of ODEs, which can constitute a serious difficulty. An alternative way to tackle this problem is to consider a nonnegative distribution function $f(x, v, t)$ describing the number density of individuals at time $t \geqslant 0$ in position $x \in \mathbb{R}^d$ with velocity $v \in \mathbb{R}^d$. The evolution of $f(x, v, t)$ is characterized by a kinetic equation which takes into account the motion of individuals due to their own velocity and the velocity changes due to the interactions with other individuals. Following [54] we consider here binary interaction Boltzmann-type and mean-field kinetic approximation of the microscopic dynamics.

1.3.1 Boltzman-Povzner kinetic approximations

In agreement with (1.3) and (1.5), we consider a microscopic binary interaction between two individuals with positions and velocities (x, v) and (y, w) according to

$$\begin{cases} v^* = (1 - \eta(H_\alpha(|x-y|, v))v + \eta H_\alpha(|x-y|, v)w, \\ \\ w^* = \eta H_\alpha(|x-y|, w)v + (1 - \eta(H_\alpha(|x-y|, w))w, \end{cases} \tag{1.12}$$

where v^*, w^* are the *post-interaction* velocities and η a parameter that measures the strength of the interaction. Analogous binary interactions can be introduced for other swarming and flocking dynamics like D'Orsogna-Bertozzi [54].

We describe the interaction of the sistem with following integro-differential equation of Boltzmann type

$$\begin{aligned} (\partial_t f + v \cdot \nabla_x f)(x, v, t) &= \frac{1}{\varepsilon} Q(f, f)(x, v, t), \\ Q(f, f) &= \int_{\mathbb{R}^{2d}} \left(\frac{1}{J} f(x, v_*, t) f(y, w_*, t) - f(x, v, t) f(y, w, t) \right) dw dy, \end{aligned} \tag{1.13}$$

where (v_*, w_*) are the *pre-interacting* velocity that generate the couple (v, w) according to (1.12), J is the Jacobian of the transformation of (v, w) to (v_*, w_*). Without visual limitation the Jacobian reads $J = (1 - 2\eta H(|x - y|))^d$. Note that, at variance with classical Boltzmann equation the interaction is non local as in Povzner kinetic model [149].

Let us introduce the time scaling

$$t \to t/\varepsilon, \qquad \eta = \lambda\varepsilon, \tag{1.14}$$

where λ is a constant and ε a small parameter. The scaling corresponds to assume that the parameter η characterizing the strength of the microscopic interactions is small, thus the frequency of interactions has to increase otherwise the collisional integral will vanish. This corresponds to large scale interaction frequencies and small interaction strengths, in agreement with a classical mean-field limit and similarly to the so-called *grazing collision limit* of the Boltzmann equation for granular gases [134].

1.3.2 Derivation of the mean-field kinetic model

First of all let us remark that the dynamic (1.12) doesn't preserve the momentum, as consequence of the velocity dependent function H_α we have

$$v^* + w^* = v + w - \eta(H_\alpha(|x-y|, w) - H_\alpha(|x-y|, v))(w - v). \tag{1.15}$$

Moreover under the assumptions $|H_\alpha(r,v)| \leqslant 1$ and $\eta \leqslant 1/2$, it is easy to prove that the support of velocity is limited by initial condition

$$v^* = (1-\eta H_\alpha(|x-y|\,,v))v + \eta H_\alpha(|x-y|\,,v)w \leqslant \max\{|v|,|w|\}. \tag{1.16}$$

Considering now the weak formulation of (1.13) the Jacobian term disappears and we get the rescaled equation

$$\begin{aligned}\frac{\partial}{\partial t}\int_{\mathbb{R}^{2d}} \phi(x,v)f(x,v,t)dvdx + \int_{\mathbb{R}^{2d}} (v\cdot\nabla\phi(x,v))f(x,v,t)dvdx = \\ \frac{1}{\varepsilon}\int_{\mathbb{R}^{4d}} (\phi(x,v^*)-\phi(x,v))f(x,v,t)f(y,w,t)dvdxdwdy,\end{aligned} \tag{1.17}$$

for $t>0$ and for all $\phi \in C_0^\infty(\mathbb{R}^{2d})$, such that

$$\lim_{t\to 0}\int_{\mathbb{R}^{2d}} \phi(x,v)f(x,v,t)dvdx = \int_{\mathbb{R}^{2d}} \phi(x,v)f_0(x,v)dvdx, \tag{1.18}$$

where $f_0(x,v)$ is the starting density.

For small values of ε we have $v^* \approx v$ thus we can consider the Taylor expansion of $\phi(x,v^*)$ around v up to the second order we obtain the following formulation to the collisional integral

$$\begin{aligned}&\frac{1}{\varepsilon}\int_{\mathbb{R}^{4d}} (\phi(x,v^*)-\phi(x,v))f(x,v,t)f(y,w,t)dvdxdwdy = \\ &\quad = \lambda\underbrace{\int_{\mathbb{R}^{4d}} (\nabla_v\phi(x,v)\cdot(w-v))H_\alpha(x,y,v)f(x,v,t)f(y,w,t)dvdxdwdy}_{:=I_1(f,f)} \\ &+\lambda^2\varepsilon\underbrace{\int_{\mathbb{R}^{4d}}\left[\sum_{i,j=1}^{d}\frac{\partial^2\phi(x,\tilde{v})}{\partial v_i^2}(w_j-v_j)^2\right](H_\alpha(x,y,v))^2 f(x,v,t)f(y,w,t)dvdxdwdy}_{:=I_2(f,f)}\end{aligned} \tag{1.19}$$

for some $\tilde{v} = \tau v + (1-\tau)v^*$, $0 \leqslant \tau \leqslant 1$. In the limit $\varepsilon \to 0$ the term $I_2(f,f)$ vanishes since the second momentum of the solution is not increasing and $H_\alpha(x,y,v) \leqslant 1$ hence [53]

$$|I_2(f,f)| \leqslant 2\|\phi(x,v)\|_{C_0^2}\int_{\mathbb{R}^{2d}} |v|^2 f_0(x,v)dxdv. \tag{1.20}$$

Thus in the limit the second-order term can be neglected and $I_1(f,f)$ constitutes an approximation of the collisional integral $Q(f,f)$, in the strong divergence form

$$I_1(f,f) = -\nabla_v\cdot\int_{\mathbb{R}^{2d}} (w-v)H_\alpha(x,y,v)f(y,w,t)f(x,v,t)dwdy, \tag{1.21}$$

or equivalently in convolution form [53]

$$I_1(f,f) = \nabla_v \cdot \{f(x,v,t)[(H_\alpha(x,y,v)\nabla_v e(v)) * f](x,v,t)\}\,, \tag{1.22}$$

where $e(v) = |v|^2/2$ and $*$ is the (x,v)-convolution. As observed in [53], the operator $I_1(f,f)$ preserves the dissipation proprieties of original Boltzmann operator.

Finally we get the mean-field kinetic equation

$$\partial_t f + v \cdot \nabla_x f = -\lambda \nabla_v \cdot (\xi(f)f) \tag{1.23}$$

$$\xi(f) = \int_{\mathbb{R}^{2d}} H_\alpha(x,y,v)(w-v)f(y,w,t)dwdy.$$

As noted in [1], the continuos kinetic model (1.23) and the microscopic one (1.3)-(1.5) are really the same when we take the discrete N-particle distribution

$$f(x,v,t) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i(t))\delta(v - v_i(t)),$$

where $\delta(\cdot)$ denotes the Dirac-delta function.

Remark 1.3.1

- *Kinetic formulation for the D'Orsogna Bertozzi et al. with perception cone can be derived in the same way and yields the mean-field model*

$$\partial_t f + v\cdot\nabla_x f + \nabla_v\cdot(S(v)f) = -\lambda\nabla_v\cdot\left(f(x,v,t)\int_{\mathbb{R}^{2d}} Z_\alpha(x,y,v)f(y,w,t)dydw\right), \tag{1.24}$$

  where $Z_\alpha(x,y,v) = (A(x,y)+R(x,y))\psi_\alpha(x,y,v)$ represents the attraction repulsion term.

- *In [164] the authors observed that a certain degree of randomness helps the coherence in the collective swarm behavior. Following [54], if we add in (1.3) a nonlinear noise term depending on function H_α and perform essentially the same derivation of the above paragraph we obtain the kinetic equation*

$$\partial_t f + v \cdot \nabla_x f = -\lambda\nabla_v \cdot (\xi(f)f) + \sigma\Delta_v[(H_\alpha * \rho)f], \tag{1.25}$$

  where $\rho = \rho(x,t)$ represent the mass of the system and $\sigma \geqslant 0$ the strength of the noise. If $H_\alpha(x,y,v) \equiv H(x,y)$, the right hand side can be written as a Fokker-Plank operator

$$\nabla_v \cdot (\sigma(H * \rho)\nabla_v f - \lambda\xi(f)f),$$

  and thus a global Maxwellian function *is a steady state solution for the equation (1.25).*

1.3.3 Alternative formulations

In this section we present some alternative formulations of the Boltzmann equation describing the binary interaction dynamics for alignment. All the formulations share the property that in the mean-field limit originate the same kinetic model (1.23).

The Boltzmann equation (1.13) has much in common with a classical Boltzmann equation for Maxwell molecules, in the sense that the collision frequency is independent of the velocity and position of individuals. An alternative Boltzmann-like kinetic approximation is obtained with the interaction operator

$$Q(f,f) = \int_{\mathbb{R}^{2d}} H_\alpha(x,y,v)\left(\frac{1}{J}f(x,v_*)f(y,w_*) - f(x,v)f(y,w)\right)dwdy, \tag{1.26}$$

where now

$$\begin{cases} v^* = (1-\eta)v + \eta w, \\ \\ w^* = \eta v + (1-\eta)w. \end{cases} \tag{1.27}$$

From the modeling viewpoint here the function H_α is interpreted as the frequency of interactions instead of the strength of the same interactions.

Clearly the two formulations (1.13) and (1.26) are not equivalent in general. It is easy to verify that formally we obtain the same mean-field limit (1.23). Note however that now the second order term in the expansion (1.19) is slightly different and reads

$$I_2(f,f) := \int_{\mathbb{R}^{4d}} \left[\sum_{i,j=1}^{d} \frac{\partial^2 \phi(x,\tilde{v})}{\partial v_i^2}(w_j - v_j)^2\right] H_\alpha(x,y,v)f(x,v,t)f(y,w,t)dvdxdwdy.$$

Since $H_\alpha > (H_\alpha)^2$, in practice we may expect a slower convergence to the mean-field dynamic for small values of ε.

Let us finally introduce some stochastic effect in the visual cone perception by defining

$$H_\alpha(x_i,x_j,v_i) = \zeta H(x_i,x_j),$$

where ζ is a random variable distributed accordingly to some $b_\alpha(\zeta,x_i,x_j,v_i) \geqslant 0$ s.t.

$$\int b_\alpha(\zeta,x_i,x_j,v_i)\,d\zeta = 1, \quad \forall\ x_i,x_j,v_i. \tag{1.28}$$

Then the collision term in the form (1.13) becomes

$$Q(f,f) = \int_{\mathbb{R}^{2d+1}} b_\alpha(\zeta,x,y,v)\left(\frac{1}{J}f(x,v_*)f(y,w_*) - f(x,v)f(y,w)\right)dwdyd\zeta, \tag{1.29}$$

whereas in the space dependent interaction frequency form (1.26) reads

$$Q(f,f) = \int_{\mathbb{R}^{2d+1}} B_\alpha(\zeta,x,y,v)\left(\frac{1}{J}f(x,v_*)f(y,w_*) - f(x,v)f(y,w)\right) dwdyd\zeta, \tag{1.30}$$

where $B_\alpha(\zeta,x,y,v) = b_\alpha(\zeta,x,y,v)H(x,y)$. Again it can be shown that thanks to (1.28) the limit asymptotic behavior $\varepsilon \to 0$ is unchanged. We omit the details.

We conclude the section reporting an example of distribution for the random variable ζ which corresponds to the stochastic analogue of (1.11)

$$\zeta = \begin{cases} 1 & \text{with probability} \quad p_1, & \text{for } y \in \Sigma(x,v,\theta), \\ 0 & \text{with probability} \quad 1-p_1, & \text{for } y \in \Sigma(x,v,\theta), \\ 1 & \text{with probability} \quad p_2, & \text{for } y \in \mathbb{R}^d \setminus \Sigma(x,v,\theta), \\ 0 & \text{with probability} \quad 1-p_2, & \text{for } y \in \mathbb{R}^d \setminus \Sigma(x,v,\theta). \end{cases}$$

1.4 Monte Carlo methods

Following [34, 77] we introduce different numerical approaches for the above kinetic equations based on Monte Carlo methods. The main idea is to approximate the dynamic by solving the Boltzmann-like models for small value of ε. We will also develop some direct Monte Carlo techniques for the limiting kinetic equation (1.23). For the sake of simplicity we describe the algorithms in the case of the collision operator (1.13), extensions to the other possible formulations presented in Section 3 are also discussed along the section. As we will see, thanks to the structure of the equations, the resulting algorithms are fully meshless.

1.4.1 Asymptotic binary interaction algorithms

As in most Monte Carlo methods for kinetic equations, see [143], the starting point is a splitting method based on evaluating in two different steps the transport and collisional part of the scaled Boltzmann-Povzner equation

$$\frac{\partial f}{\partial t} = -v \cdot \nabla_x f \tag{T}$$

$$\frac{\partial f}{\partial t} = \frac{1}{\varepsilon} Q_\varepsilon(f,f) \tag{C}$$

where we used the notation $Q_\varepsilon(f,f)$ to denote the scaled Boltzmann operator (1.13). We emphasize that the solution to the collision step for small values of ε has very little in common with the classical fluid-limit of the Boltzmann equation. Here in fact the whole collision process depends on space and on the small scaling parameter

ε. In particular, in the small ε limit the solution is expected to converge towards the solution of the mean-field model (1.23).

By decomposing the collisional operator in equation (C) in its gain and loss parts we can rewrite the collision step as

$$\frac{\partial f}{\partial t} = \frac{1}{\varepsilon}\left[Q_{\varepsilon}^{+}(f,f) - \rho f\right], \tag{1.31}$$

where $\rho > 0$ represent the total mass and Q_{ε}^{+} the *gain* part of the collisional operator. Without loss of generality in the sequel we assume that

$$\rho = \int_{\mathbb{R}^{2d}} f(x,v,t)dxdv = 1.$$

In order to solve the trasport step we use the exact free flow of the sample particles $(x_i(t), v_i(t))$ in a time interval Δt

$$x_i(t+\Delta t) = x_i(t) + v_i(t)\Delta t, \tag{1.32}$$

and thus describe the different schemes used for the interaction process in the form (1.31).

A Nanbu-like asymptotic method

Let us now consider a time interval $[0,T]$ discretized in n_{tot} intervals of size Δt. We denote by f^n the approximation of $f(x,v,n\Delta t)$.

Thus the forward Euler scheme writes

$$f^{n+1} = \left(1 - \frac{\Delta t}{\varepsilon}\right) f^n + \frac{\Delta t}{\varepsilon} Q_{\varepsilon}^{+}(f^n, f^n), \tag{1.33}$$

where since f^n is a probability density, thanks to mass conservation, also $Q_{\varepsilon}^{+}(f^n, f^n)$ is a probability density. Under the restriction $\Delta t \leqslant \varepsilon$ then also f^{n+1} is a probability density, since it is a convex combination of probability densities.

From a Monte Carlo point of view equation (1.33) can be interpreted as follows: an individual with velocity v at position x will not interact with other individuals with probability $1 - \Delta t/\varepsilon$ and it will interact with others with probability $\Delta t/\varepsilon$ according to the interaction law stated by $Q_{\varepsilon}^{+}(f^n, f^n)$. Since we aim at small values of ε the natural choice as in [34] is to take $\Delta t = \varepsilon$. The major difference compare to standard Nanbu algorithm here is the way particles are sampled from $Q_{\varepsilon}^{+}(f^n, f^n)$ which does not require the introduction of a space grid. A simple algorithm for the solution of (1.33) in a time interval $[0,T]$, $T = n_{tot}\Delta t$, $\Delta t = \varepsilon$ is sketched in the sequel.

Algorithm 1.4.1 (Asymptotic Nanbu I)

```
1. Given $N$ samples $(x_k^0, v_k^0)$, with $k = 1, \dots, N$ from the initial distribution
   $f_0(x, v)$;

2. for $n = 0$ to $n_{tot} - 1$

        for $i = 1$ to $N$;
        (a) select an index $j$ uniformly among all possible individuals $(x_k^n, v_k^n)$
            except $i$;
        (b) evaluate $H_\alpha(|x_i^n - x_j^n|, v_i^n)$;
        (c) compute the velocity change $v_i^*$ using the first relation in (1.12) with
            $\eta = \varepsilon$;
        (d) set $(x_i^{n+1}, v_i^{n+1}) = (x_i^n, v_i^*)$.
        end for

   end for
```

Next we show how the method extends to the case of collision operator of the type (1.26). In this case an acceptance-rejection strategy is used to select interacting individuals since the forward Euler scheme reads

$$f^{n+1} = \left(1 - \frac{\Delta t}{\varepsilon}\right) f^n + \frac{\Delta t}{\varepsilon} P_\varepsilon^+(f^n, f^n), \tag{1.34}$$

where $P_\varepsilon^+(f^n, f^n) = Q_\varepsilon(f, f) + f \geqslant 0$ is again a probability density.

Now using the fact that $H_\alpha \leqslant 1$ we can adapt the classical acceptance-rejection technique [143] to get the following method for (1.34) with $\Delta t = \varepsilon$

Algorithm 1.4.2 (Asymptotic Nanbu II)

```
In Algorithm 1.4.1 make the following change

(c) if $H_\alpha(|x_i^n - x_j^n|, v_i^n) > \xi$, $\xi$ uniform in $[0, 1]$ then compute the velocity
    change $v_i^*$ using the first relation in (1.27) with $\eta = \varepsilon$;
(d) set $(x_i^{n+1}, v_i^{n+1}) = (x_i^n, v_i^*)$ if the individual has changed its velocity, otherwise set $(x_i^{n+1}, v_i^{n+1}) = (x_i^n, v_i^n)$.
```

Note that in this version two individuals interact always with the same strength in the velocity change but with a different probability related to their distance. As a result the total number of interactions depend on the distribution of individuals and on average is equal to $\bar{H}_\alpha N < N$ where

$$\bar{H}_\alpha = \frac{1}{N^2} \sum_{i,j=1}^{N} H_\alpha(x_i, x_j, v_i).$$

Thus the method computes less interactions then the one described in Algorithm 1.4.1. In fact, in regions where individuals are scattered very few interactions will be effectively computed by the method. The efficiency of the method can be further improved if one is able to find an easy invertible function $1 \geqslant W_\alpha(x_i, x_j, v_i) \geqslant H_\alpha(x_i, x_j, v_i)$ or is capable to compute directly the inverse of $H_\alpha(x_i, x_j, v_i)$. We refer to [143] for further details on these sampling techniques.

A symmetric version of the previous algorithms which preserves at a microscopic level other interaction invariants, like momentum in standard Cucker-Smale model, is obtained as follows

Algorithm 1.4.3 (Asymptotic symmetric Nanbu)

```
1. Given $N$ samples $(x_k^0, v_k^0)$, with $k = 1, \dots, N$ from the initial distribution $f_0(x, v)$;
2. for $n = 0$ to $n_{tot} - 1$
    (a) set $N_c = Iround(N/2)$;
    (b) select $N_c$ random pairs $(i, j)$ uniformly without repetition among all possible pairs of individuals at time level $n$.
    (c) evaluate $H_\alpha(|x_i^n - x_j^n|, v_i^n)$ and $H_\alpha(|x_i^n - x_j^n|, v_j^n)$;
    (d) For Algorithm 1.4.1: compute the velocity changes $v_i^*$, $v_j^*$ for each pair $(i, j)$ using relations (1.12) with $\eta = \varepsilon$;
    (d) For Algorithm 1.4.2:
        i. if $H_\alpha(|x_i^n - x_j^n|, v_i^n) > \xi_i$ $\xi_i$ uniform in $[0, 1]$ then compute the velocity change $v_i^*$ for each pair $(i, j)$ using the first relation in (1.27) with $\eta = \varepsilon$;
        ii. if $H_\alpha(|x_i^n - x_j^n|, v_j^n) > \xi_j$ $\xi_j$ uniform in $[0, 1]$ then compute the velocity change $v_j^*$ for each pair $(i, j)$ using the second relation in (1.27) with $\eta = \varepsilon$;
    (e) set $(x_i^{n+1}, v_i^{n+1}) = (x_i^n, v_i^*)$, $(x_j^{n+1}, v_j^{n+1}) = (x_j^n, v_j^*)$ for all the individuals that changed their velocity,
    (f) $(x_h^{n+1}, v_h^{n+1}) = (x_h^n, v_h^n)$ for all the remaining individuals.
end for
```

The function $Iround(\cdot)$ denotes the integer stochastic rounding defined as

$$Iround(x) = \begin{cases} [x] + 1, & \xi < x - [x], \\ [x], & \text{elsewhere} \end{cases}$$

where ξ is a uniform $[0, 1]$ random number and $[\cdot]$ is the integer part.

A Bird-like asymptotic method

The most popular Monte Carlo approach to solve the collision step in Boltzmann-like equations is due to Bird [32]. The major differences are that the method simulate the time continuous equation and that individuals are allowed to interact more then once in a single time step. As a result the method achieves a higher time accuracy [143].

Here we describe the algorithm for the collision operator described by (1.13). The method is based on the observation that the interaction time is a random variable exponentially distributed. Thus for N individuals one introduces a local random time counter given by

$$\Delta t_c(\xi) = -\frac{\ln(\xi)\varepsilon}{N}, \tag{1.35}$$

with ξ a random variable uniformly distributed in $[0, 1]$.

A simpler version of the method is based on a constant time counter Δt_c corresponding to the average time between interactions. In fact, in a time interval $[0, T]$ we have

$$\Delta t_c = \frac{T}{N_c} = \frac{\varepsilon}{N}, \tag{1.36}$$

since $N_c = NT/\varepsilon$ is the number of average interactions in the time interval. Of course taking time averages the two formulations (1.35) and (1.36) are equivalent.

From the above considerations, using the symmetric formulation and the time counter $\Delta t_c = 2\varepsilon/N$, we obtain the following method in a time interval $[0, T]$, $T = N_{tot}\Delta t_c$

Algorithm 1.4.4 (Asymptotic Bird I)

```
1. Given $N$ samples $(x_k, v_k)$, with $k = 1, \dots, N$ from the initial distribution $f_0(x, v)$
2. for $n = 0$ to $N_{tot} - 1$
```

```
    (a) select a random pair $(i, j)$ uniformly among all possible pairs;
    (b) evaluate $H_\alpha(|x_i - x_j|, v_i)$ and $H_\alpha(|x_i - x_j|, v_j)$;
    (c) compute the velocity changes $v_i^*$, $v_j^*$ using relations (1.12) with $\eta = \varepsilon$;
    (d) set $v_i = v_i^*$ and $v_j = v_j^*$;
end for
```

Note that in the above formulation the method has much in common with Algorithm 1.4.3 except for the fact that multiple interactions are allowed during the dynamic (no need to tag particles with respect to time level) and that the local time stepping is related to the number of individuals. As a result in the limit of large numbers of individuals the method converges towards the time continuous Boltzmann equation (1.13) and not to its time discrete counterpart (1.33), as it happens for Nanbu formulation. Since in Algorithm 1.4.3 we have $n_{tot} = N_{tot}/N_c$, the computational cost of the methods is the same.

Similarly Bird's approach can be extended to collision operator in the form (1.26) by introducing the following changes

Algorithm 1.4.5 (Asymptotic Bird II)

```
In Algorithm 1.4.4 make the following change
    (c) * if $H_\alpha(|x_i^n - x_j^n|, v_i^n) > \xi_i$ $\xi_i$ uniform in $[0, 1]$ then compute the velocity
          change $v_i^*$ using the first relation in (1.27) with $\eta = \varepsilon$;
        * if $H_\alpha(|x_i^n - x_j^n|, v_j^n) > \xi_j$ $\xi_j$ uniform in $[0, 1]$ then compute the velocity
          change $v_j^*$ using the second relation in (1.27) with $\eta = \varepsilon$;
```

Finally we sketch the algorithm to implement the stochastic perception cone present in (1.29) and (1.30), that can be easily introduced in all the previous algorithms.

Algorithm 1.4.6 (Interaction with stochastic perception cone)

```
• if $x_j \in \Sigma(x_i, v_i, \theta)$
    – with probability $p_1$ perform the interaction between $i$ and $j$ and compute
      $v_i^*$
  else
```

```
    – with probability $p_2$ perform the interaction between $i$ and $j$ and compute
      $v_i^*$

• if $x_i \in \Sigma(x_j, v_j, \theta)$
    – with probability $p_1$ perform the interaction between $i$ and $j$ and compute
      $v_j^*$
  else
    – with probability $p_2$ perform the interaction between $i$ and $j$ and compute
      $v_j^*$
```

Note that this reduces further the total number of interactions in the algorithms just described. In contrast, for the deterministic case we simply change the relative interaction strengths using respectively $\eta = p_1\varepsilon$ and $\eta = p_2\varepsilon$ in the binary interaction rules.

1.4.2 Mean-field interaction algorithms

Let us finally tackle directly the limiting mean field equation. The interaction step now corresponds to solve

$$\partial_t f = -\nabla_v \cdot \left(f \int_{\mathbb{R}^{2d}} H_\alpha(x,y,v)(v-w) f(y,w,t) dw dy \right).$$

As already observed, in a particle setting this corresponds to compute the original $O(N^2)$ dynamic. We can reduce the computational cost using a Monte Carlo evaluation of the summation term as described in the following simple algorithm.

Algorithm 1.4.7 (Mean Field Monte Carlo)

```
1. Given $N$ samples $v_k^0$, with $k = 1, \dots, N$ computed from the initial distribution
   $f_0(x, v)$ and $M \leqslant N$;
2. for $n = 0$ to $n_{tot} - 1$
    (a) for $i = 1$ to $N$
    (b) sample $M$ particles $j_1, \dots, j_M$ uniformly without repetition among all particles;
    (c) compute
        $\bar{H}^n_\alpha(x_i) = \frac{1}{M}\sum_{k=1}^{M} H_\alpha(x_i, x_{j_k}, v_i^n), \quad \bar{v}_i^n = \frac{1}{M}\sum_{k=1}^{M} \frac{H_\alpha(x_i^n, x_{j_k}^n, v_i^n)}{\bar{H}^n_\alpha(x_i)} v_{j_k},$
    (d) compute the velocity change
        $v_i^{n+1} = v_i^n(1 - \Delta t \bar{H}^n_\alpha(x_i)) + \Delta t \bar{H}^n_\alpha(x_i) \bar{v}_i^n.$
  end for
end for
```

The overall cost of the above simple algorithm is $O(MN)$, clearly for $M = N$ we obtain the explicit Euler scheme for the original N particle system. In this formulation the method is closely related to asymptotic Nanbu's Algorithm 1.4.1. It is easy to verify that taking $M = 1$ leads exactly to the same numerical method. On the other hand for $M > 1$ the above algorithm can be interpreted as an averaged asymptotic Nanbu method over M runs since we can rewrite point (d) as

$$v_i^{n+1} = \frac{1}{M}\sum_{k=1}^{M} \left[\left(1 - \Delta t H_\alpha(x_i^n, x_{j_k}^n, v_i^n)\right) v_i^n + \Delta t H_\alpha(x_i^n, x_{j_k}^n, v_i^n) v_{j_k}^n\right], \qquad i = 1, \dots, N.$$

The only difference is that averaging the result of Algorithm 1.4.1 does not guarantee the absence of repetitions in the choice of the indexes $j_1, \dots, j_M$. Thus the choice $\Delta t = \varepsilon$ in Algorithm 1.4.1 originates a numerical method consistent with the limiting mean-field kinetic equation. Following this description we can construct other Monte Carlo methods for the mean field limit taking suitable averaged versions of the corresponding algorithms for the Boltzmann models. Here we omit for brevity the details.

Remark 1.4.1

- *In Algorithm 1.4.7 the size of Δt can be taken larger then the corresponding $\Delta t = \varepsilon$ in Algorithm 1.4.1. However, as we just discussed, since large values of Δt in the mean-field algorithm are essentially equivalent to large values of ε in the Boltzmann algorithms we don't expect any computational advantage by choosing larger values of Δt in Algorithm 1.4.7.*

- *We remark that changing the time discretization method from Explicit Euler in (1.33) and (1.34) to other methods, like semi-implicit methods or method designed for the fluid-limit, permits to avoid the stability restriction $\Delta t \leqslant \varepsilon$. Even this approach however does not lead to any computational improvement since a strong deterioration in the accuracy of the solution is observed when $\Delta t > \varepsilon$. Here we don't explore further this direction.*

1.5 Numerical Tests

In this section we first compare accuracy and computational cost of some of the different methods and then illustrate their performance on different two-dimensional and three-dimensional numerical examples. We use the following notations: $ANMC$ (Algorithm 1.4.3), $ABMC$ (Algorithm 1.4.4), and $MFMC_M$ (Algorithm 1.4.7 for a given M).

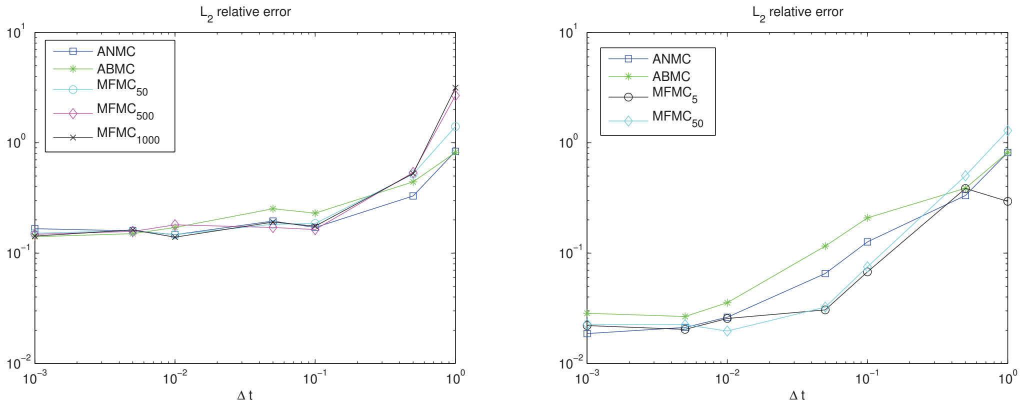


Figure 1.3: Relative errors in the L_2 norm at $T = 1$ for the different methods as a function of $\Delta t = \varepsilon$. On the left the error is computed with $N = 1000$ particles, on the right the same test is performed with $N = 50000$ particles.

1.5.1 Accuracy considerations

Here we compare the accuracy of the different algorithms studied for a simple space homogeneous situation. In fact, since the algorithms differ only in the binary interaction dynamic the homogeneous step is the natural setting for comparing the various approaches in term of accuracy.

We consider the standard Cucker-Smale dynamic. Since we do not have any space dependence we assume $H(|x_i - x_j|) \equiv 1$ for each i, j. Thus there is no difference in this test case between formulations (1.13) and (1.26) and the relative simulation schemes.

We take $N = 50000$ individuals and at the initial time the velocity is distributed as the sum of two gaussian

$$f_0(v) = \frac{1}{\sqrt{2\pi}\sigma_v}\left(e^{-\dfrac{(v+v_0)^2}{2\sigma_v^2}} + e^{-\dfrac{(v-v_0)^2}{2\sigma_v^2}}\right),$$

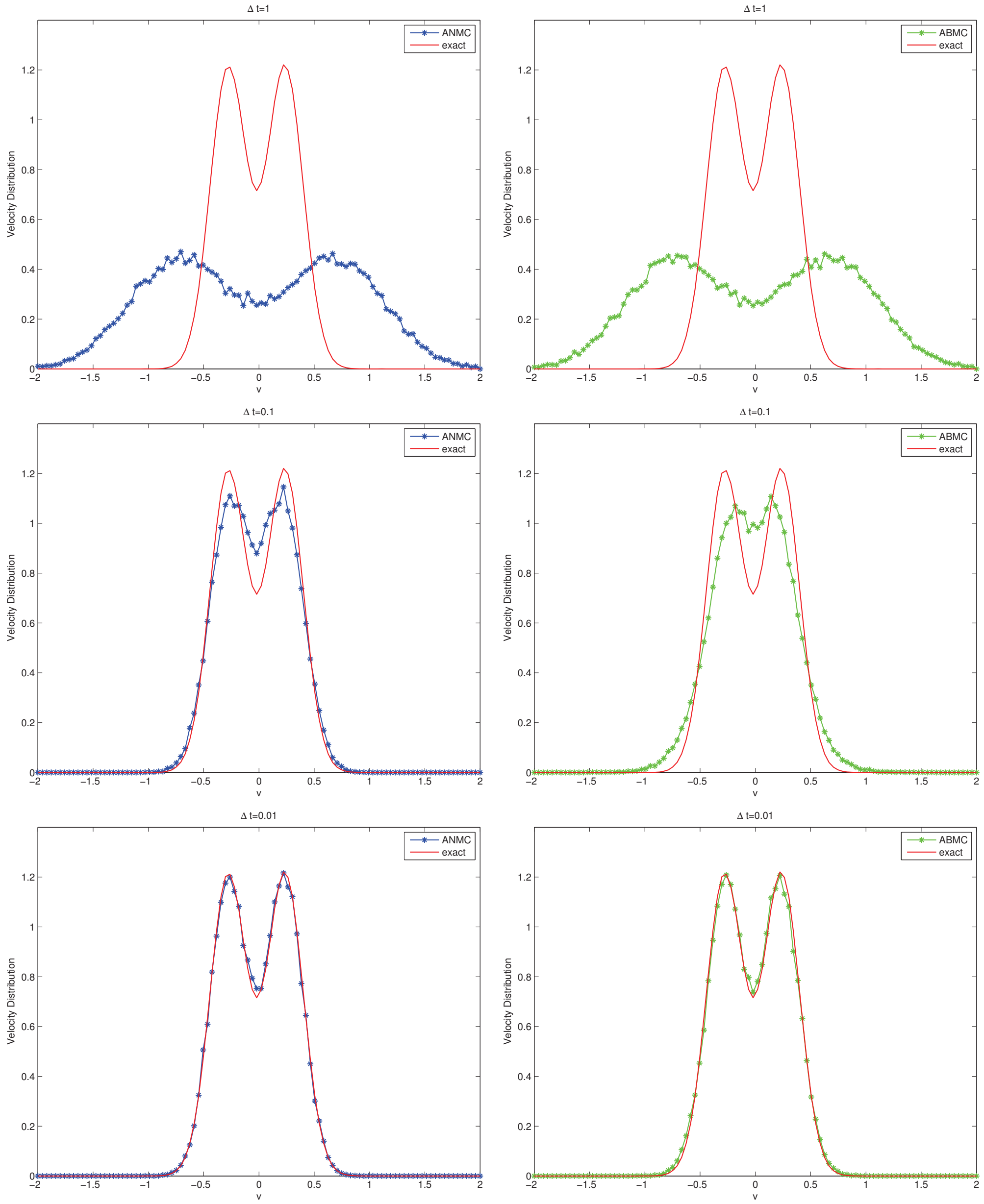


Figure 1.4: Convergence to the exact solution (continuous line) of the velocity profiles calculated with $ANMC$ (left) and $ABMC$ (right) algorithms. From the top to bottom, $\Delta t = \varepsilon$ with $\varepsilon = 1, 0.1, 0.01$.

with $v_0 = 0.7$, $\sigma_v = \sqrt{0.2}$.

The results obtained for $ANMC$ and $ABMC$ with $\varepsilon = 1, 0.1, 0.01$ at time $T = 1$ are reported in Figure 1.4. The reference solution is computed using the microscopic

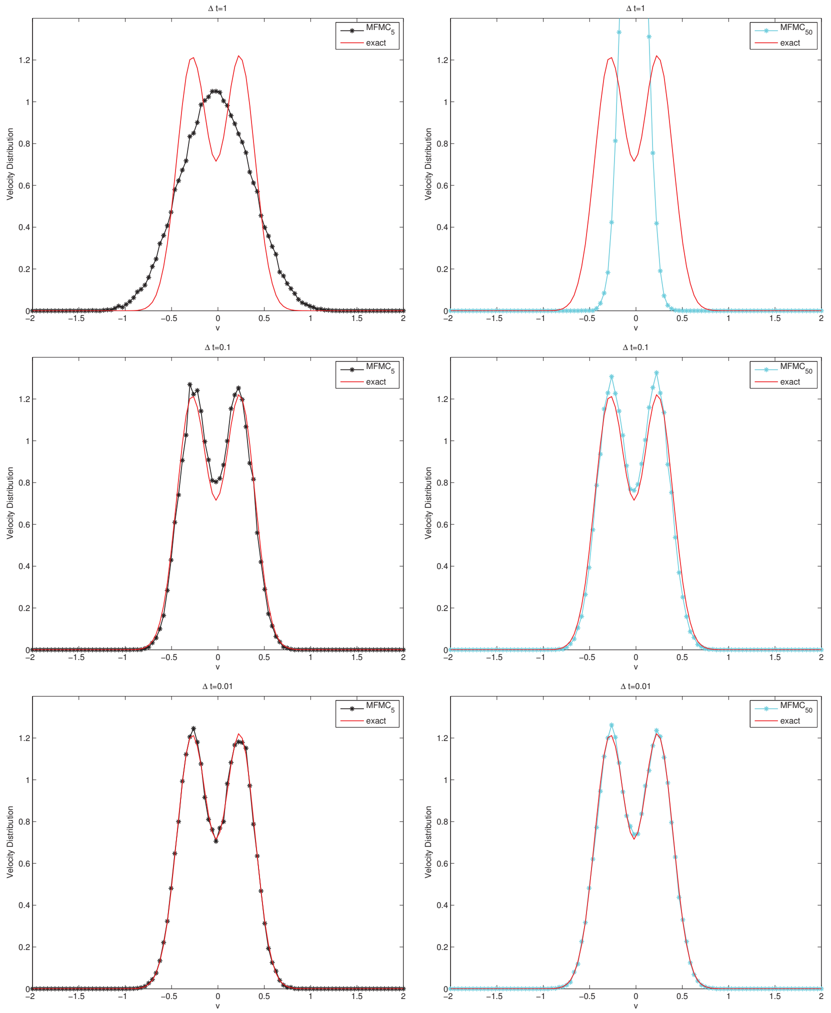


Figure 1.5: $MFMC_M$ algorithms compared with $ANMC$ method, at different time steps. From the top $\Delta t = 1$, $\Delta t = 0.1$ and $\Delta t = 0.01$. On the left column $M = 5$ on the left $M = 50$.

model which in this simple situation can be solved exactly and gives

$$v_i(t) = v_i(0)e^{-t} + (1 - e^{-t})\bar{v}, \qquad \bar{v} = \frac{1}{N}\sum_{j=1}^{n} v_j(0).$$

As expected convergence towards the exact solution is observed for both methods. In particular for $\varepsilon = 0.01$ the results are in good agreement with the direct solution of the microscopic model.

Next in Figure 1.5 we compare the behavior of the $MFMC_M$ method with $ANMC$ for the same values of the time step. A considerable difference is observed for large values of Δt and both methods are poorly accurate. On the other hand for smaller values of Δt they both converge towards the reference solution.

Finally in Figure 1.3 we report the L_2-norm of the error for $ANMC$, $ABMC$ and $MFMC_M$ for various M as a function of $\Delta t = \varepsilon$. We compare the convergence of the schemes with different number of particles $N = 1000$ and $N = 50000$. Note that in both cases the convergence rate of the schemes is rather close and for $\varepsilon = \Delta t < t^*$, the statistical error dominates the time error so that we observe a saturation effect, where $t^* \approx 0.1$ for $N = 1000$ and $t^* \approx 0.01$ for $N = 50000$.

1.5.2 Computational considerations and 1D simulations

Next we want to compare the computational cost of the different binary interaction methods for solving the kinetic equation (1.23), when compared to the direct numerical solution of the original system (2.1).

We consider the same one-dimensional test problem as in [54] for the Cucker-Smale dynamic. The initial distribution is given by

$$f_0(x,v) = \frac{1}{2\pi\sigma_x\sigma_v} e^{\dfrac{-x^2}{2\sigma_x^2}} \left(e^{\dfrac{-(v+v_0)^2}{2\sigma_v^2}} + e^{\dfrac{-(v-v_0)^2}{2\sigma_v^2}} \right),$$

for $v_0 = 2.5$, $\sigma_v = \sqrt{0.1}$ and $\sigma_x = \sqrt{2}$.

The computational time for the different methods at $T = 1$ using $\varepsilon = 0.01$ and different number of individuals is reported in Table 1.1. Simulations have been performed on a Intel Core $I7$ dual-core machine using Matlab. The $O(N)$ cost of $ANMC$ and $ABMC$ is evident. The same results are also reported in Figure 1.6 which shows the linear growth of the various Monte Carlo algorithms.

N	10^3	10^4	10^5	10^6
$ANMC$	0.02 s	0.23 s	2.82 s	3.83×10^1 s
$ABMC$	0.02 s	0.21 s	2.20 s	3.14×10^1 s
$MFMC_{50}$	0.05 s	0.41 s	4.26 s	4.44×10^1 s
$MFMC_{500}$	0.14 s	1.58 s	1.33×10^1 s	3.14×10^2 s
$MFMC_{5000}$	5.00 s	5.20×10^1 s	1.71×10^3 s	4.49×10^4 s

Table 1.1: Computational times for the different methods with various values of N. The final time is $T = 1$ and the scaling factor $\varepsilon = \Delta t$ is fixed at 10^{-2}.

Finally we report the phase-space plots of the previous 1D problem obtained using the perception cone (1.11). Clearly the parameter θ has no meaning in the one-dimensional case, so that the perception limitation concerns only the capability to detect other individuals on the left and on the right over the line. We compare

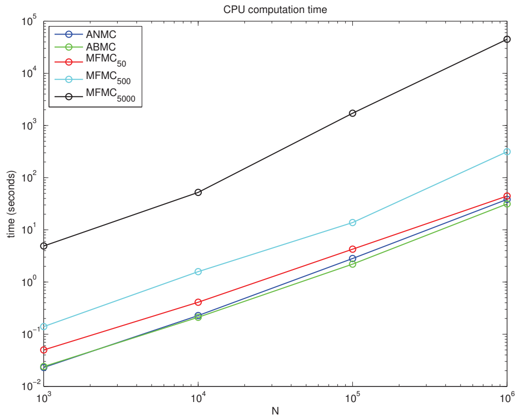


Figure 1.6: Comparison of the computational times for the different methods. For each method the time step is equal to $\Delta t = 0.01$.

the evolution in the phase space of two different cases: the classical Cucker-Smale model (non visual limitation $p_1 = p_2 = 1$) and the weighted visual cone with $p_1 = 1$ and $p_2 = 0.5$. The results are reported in Figure 1.7.

The simulations have been computed using $ABMC$ with $\Delta t = \varepsilon = 0.01$. The number N of individuals is $N = 50000$, with $\gamma = 0.05$, that is unconditional flock condition. The phase space representation is obtained using a space-velocity grid of 100×150 cells over the box $[-15, 15] \times [-10, 10]$. The results are in good agreement with the one presented in [54]. Note how perception limitations reduce the flocking tendency of the group of individuals by creating two different flocks moving towards left and right respectively.

1.5.3 2D Simulations

Cucker-Smale dynamics A generalization of the previous test in two-dimension is obtained by considering a group of N individuals with position $(x, y) \in \mathbb{R}^2$, initially distributed as

$$f_0(x, y, v_x, v_y) = g_0(x, y) h_0(v_x, v_y),$$

where

$$g_0(x,y) = \frac{1}{2\pi\sigma^2}\exp\{-(x^2+y^2)/2\sigma^2\}, \qquad h(r) = \frac{1}{2\pi\nu^2}\left(e^{\frac{-(r+v_0)^2}{2\nu^2}} + e^{\frac{-(r-v_0)^2}{2\nu^2}} \right),$$

with $r = |(v_x, v_y)|$, $v_0 = 10$, $\sigma = \sqrt{2}$ and $\nu = \sqrt{0.1}$. In the following simulations we use $N = 100000$ particles and the limited perception cone defined by (1.11).

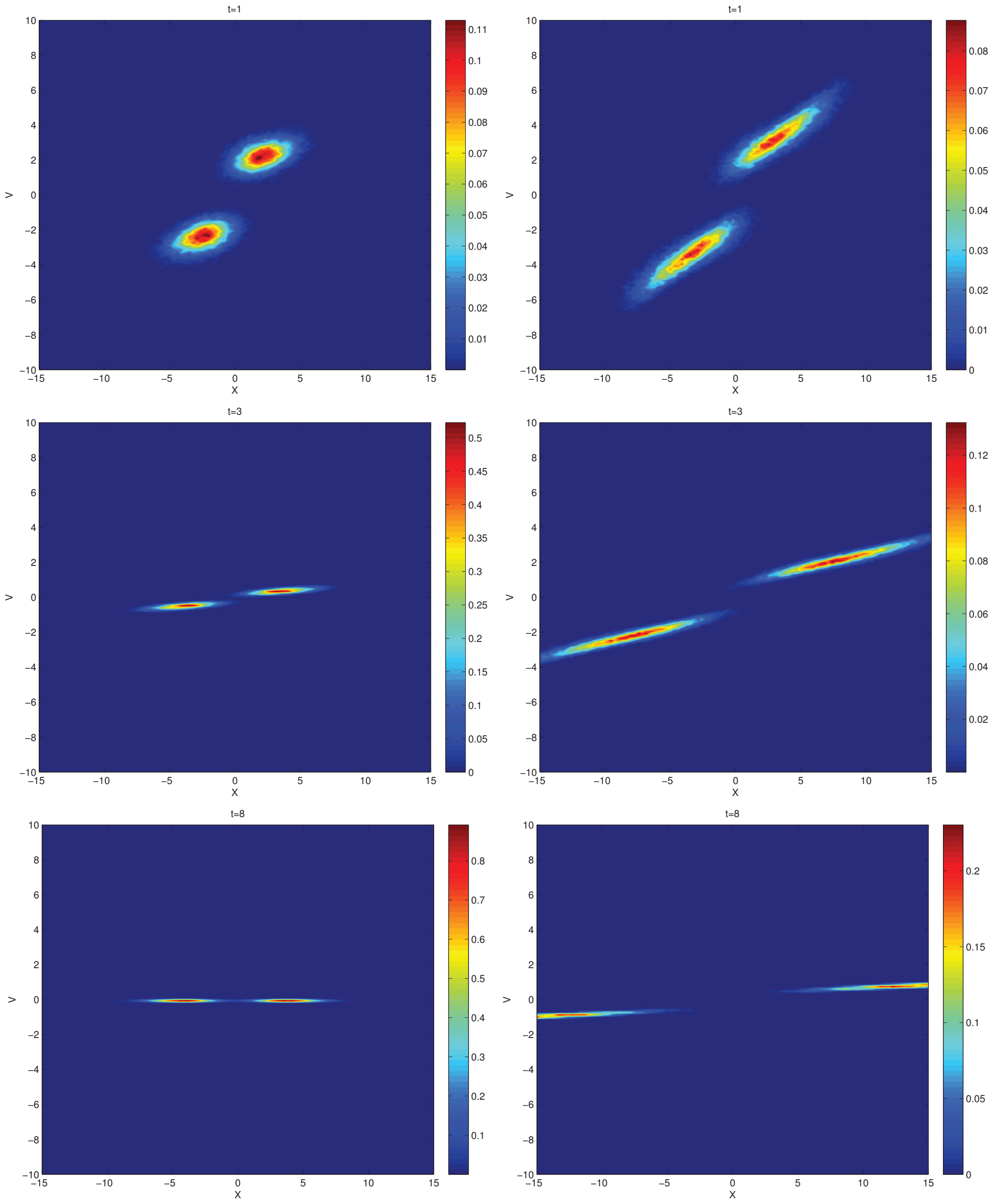


Figure 1.7: 1D Cucker-Smale flocking in the phase space. On the left without perception limitations, on the right with a perception bound characterized by $p_1 = 1$ and $p_2 = 0.5$.

We compare the evolution of the space density for different choices of the perception parameters and at different times. In the test case considered the parameters

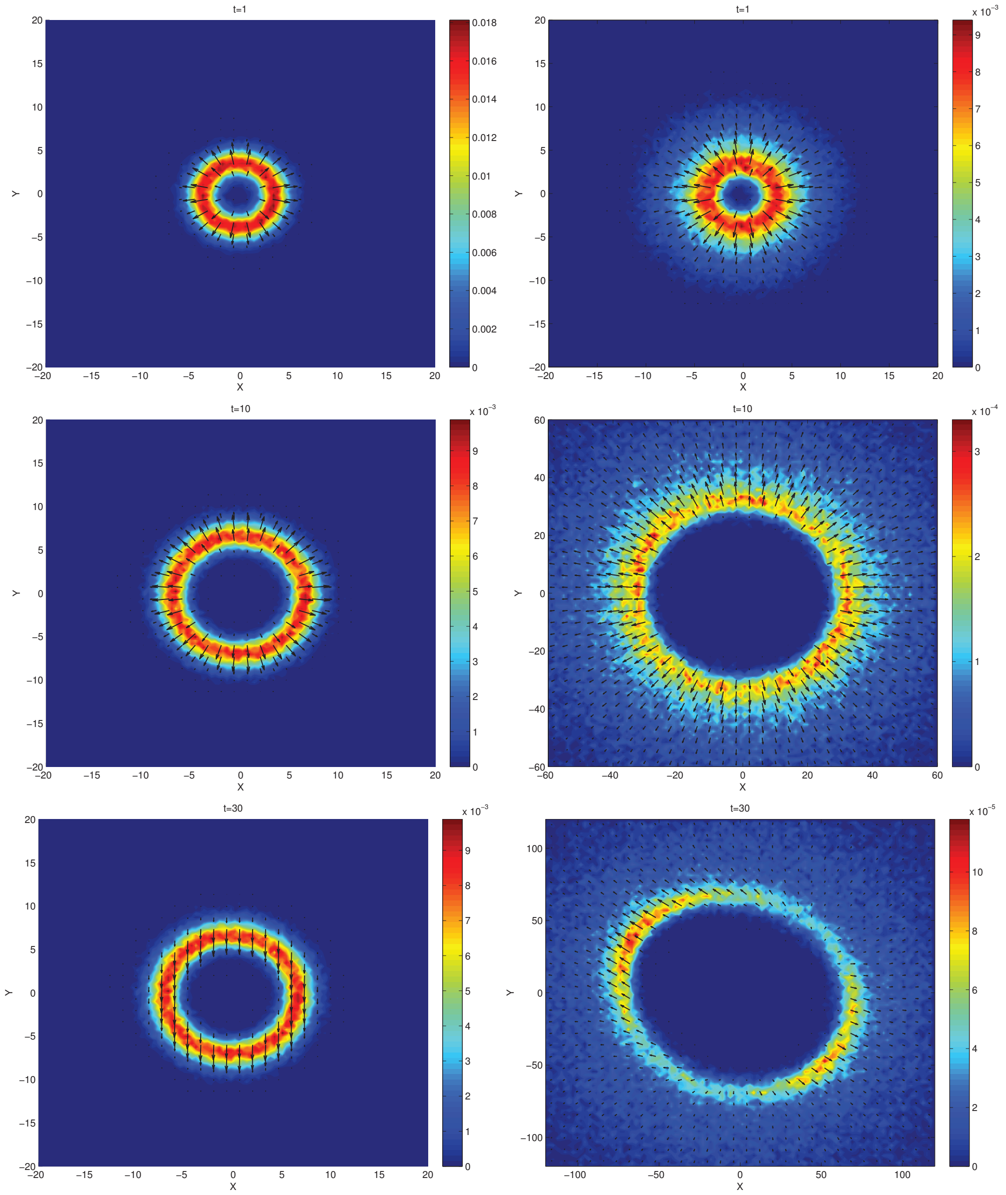


Figure 1.8: 2D Cucker-Smale flocking. On the left without perception limitations, on the right with perception cone with $\theta = 4/3\pi$, $p_1 = 1$ and $p_2 = 0.04$.

for the perception cone are $\theta = 4/3\pi$, $p_1 = 1$ and $p_2 = 0.04$.

To reconstruct the probability density function in the space we use a 100×100 grid on $[-20, 20] \times [-20, 20]$. In each figure we also add the velocity flux to illustrate the flock direction. We report the results computed with *ABMC* method and $\Delta t = \varepsilon = 0.01$, similar results are obtained with the other stochastic binary algorithms.

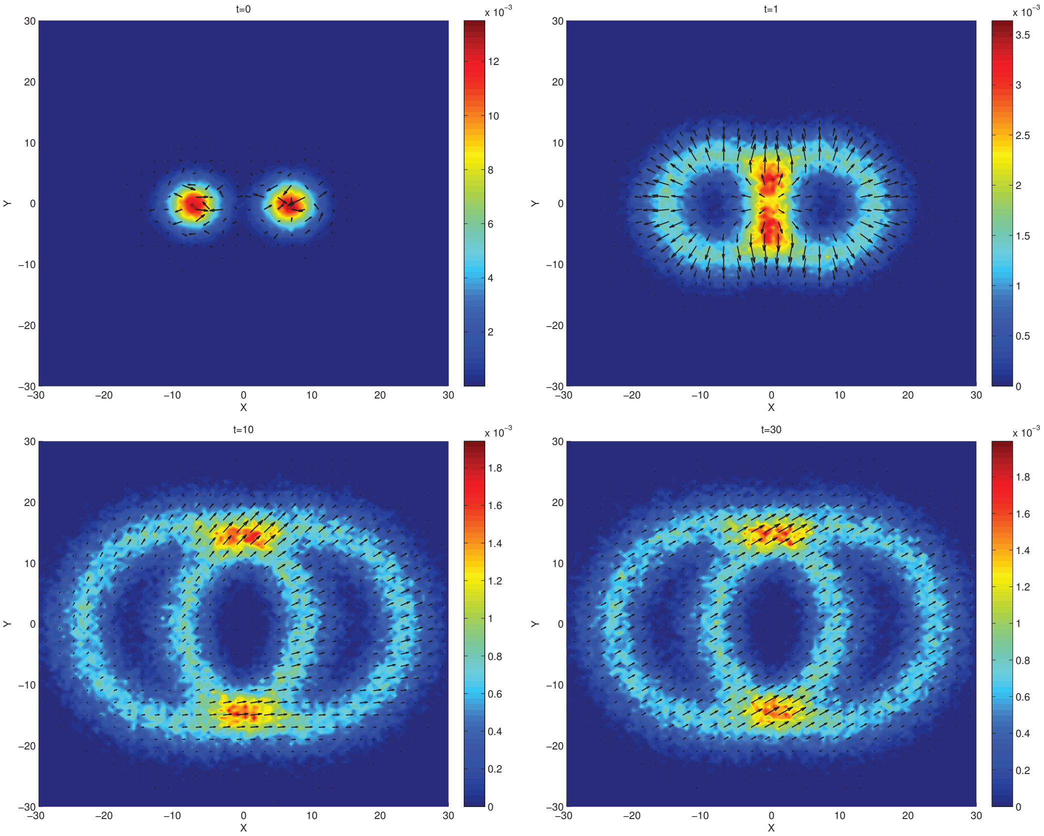


Figure 1.9: 2D Cucker-Smale dynamics. Spatial density of two flock that merge together.

At $T = 30$ the final flocking structure is reached. It is remarkable that in absence of perception limitations we obtain a perfect circular ring moving at constant speed.

In contrast when we introduce limitations the flocking behavior is less evident and the groups splits in two flocks moving in opposite directions. Finally we can also modify the previous example to create more complex patterns, but with the same basic structure.

The initial distribution now is given by

$$g_0(x, y, v_x, v_y) = f_0(x + m, y, v_x, v_y) + f_0(x - m, y, v_x, v_y),$$

where f_0 is defined as before, and $m = 7$. We report the results obtained in absence of perception cone. The final flocking state is reached at $t \approx 30$ (see Figure 1.9).

D'Orsogna, Bertozzi model et al. dynamic Next we want to simulate the D'Orsogna Bertozzi model et al. model with the aim to reproduce the typical mill

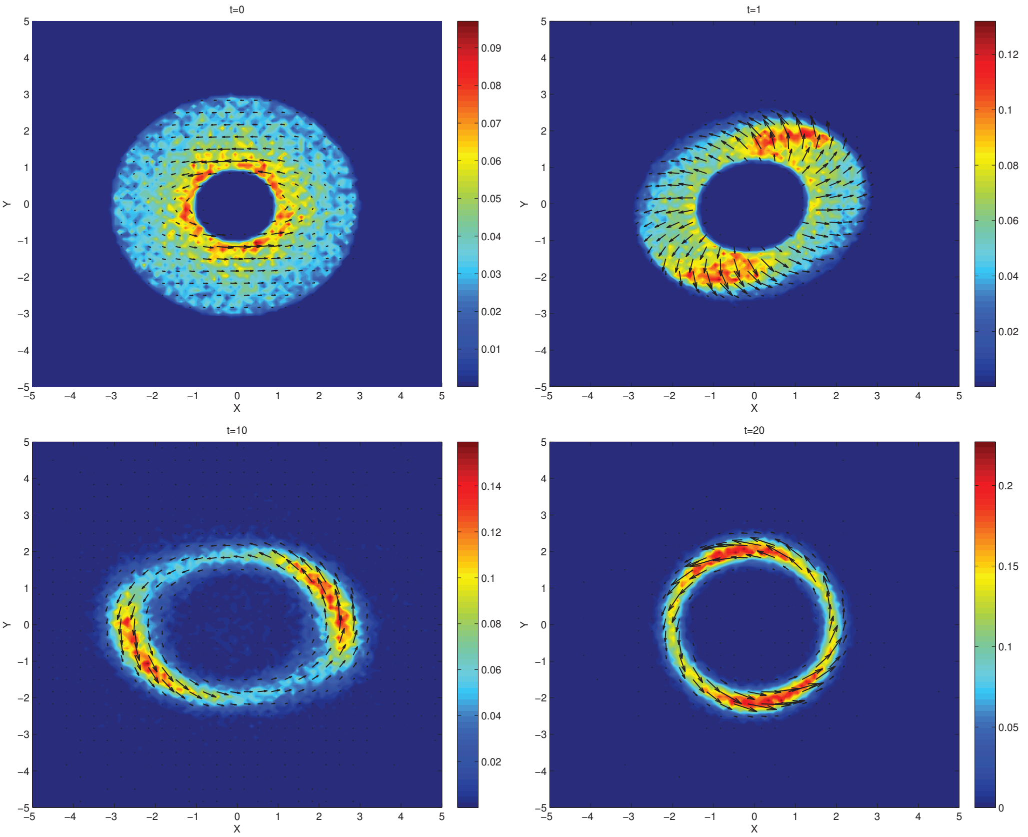


Figure 1.10: 2D Mill in D'Orsogna-Bertozzi et al. model at various times. Parameters in the attraction-repulsion potential are such that $C = C_R/C_A = 30$, $l = l_R/l_A = 0.3$, $\alpha = 0.7$, $\beta = 0.05$. Final configuration is reached after $t = 20$.

dynamics as in [50, 51, 1] but using the Boltzmann kinetic approximation.

Mills and double mills are typical emergence phenomena in school of fishes and flock of birds which travel in a compact circular motion, for example, in order to protect themselves from predator attacks. At first we work in the twodimensional space taking into account $N = 100000$ individuals. According to the interaction described in (1.6), we consider the long-range attraction and short-range repulsion.

In Figure 1.10 the initial data is uniformly distributed on a twodimensional torus, with a circular motion. The evolution shows how the attraction and repulsion forces stretch the mill reaching after $t = 20$ a condition of equilibrium in a stable circular motion as a single mill.

In Figure 1.11, we instead consider the following initial data

$$f_0(x,y,v_x,v_y) = \frac{1}{4\pi^2\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \left(e^{-\frac{(v_x+v_0)^2}{2}} + e^{-\frac{(v_x-v_0)^2}{2}} \right),$$

where $\sigma = \sqrt{2}$ and $v_0 = 0.5$. Thus density in space is a normal distribution centered in zero and velocity distribution has two main directions left and right. The evolution computed with $ABMC$ and $\Delta t = \varepsilon$ shows that equilibrium is reached after $t = 30$ in a stable double mill formation.

1.5.4 3D simulations

Finally we present some three dimensional simulations for the models taking into account the different effects of the thee zone dynamic. All the simulations have been performed with $ABMC$ and $\Delta t = \varepsilon = 0.01$.

Bertozzi-D'Orsogna et al. model In Figure 1.12 we consider the tridimensional extension of the previous simulation for the Bertozzi-D'Orsogna model et al. Initial data is uniformly distributed in space on a 3D-torus, and initial velocity is described by a circular motion in the (x, y) components in z direction initial velocity has no influence.

We present the evolution of the swarm mass density and the vectorial field. The equilibrium reached after $t = 80$ is a ball-shaped flock with mass concentrated on the border and empty zones in the middle, that is the typical configuration observed for a mill of a fish school.

Simulation is made taking in account N=200000 particles, and reconstructing the probability density function in the space we use a 3D grid with $\Delta x \times \Delta y \times \Delta z = 100 \times 100 \times 100$.

Roosting Force Accordingly to the work [51] we introduce in the D'Orsogna-Bertozzi model a roosting force term. The term expresses essentially the tendency of a flock or a school of fishes to stay around a certain zone. Such zones usually are of food interest or where birds settle their nests. Different approaches can be used to model this biological behavior, see for example [113, 19].

Mathematically speaking such term can be described by the introduction of a force term of the type

$$F_{roost} = -\left[v_i^{\perp} \cdot \nabla\phi(x_i) \right] v_i^{\perp}. \tag{1.37}$$

Such force gives the individuals a tendency to move towards the origin, for a suitable function ϕ. Here ϕ, called *roosting potential* is a function $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}$. In the

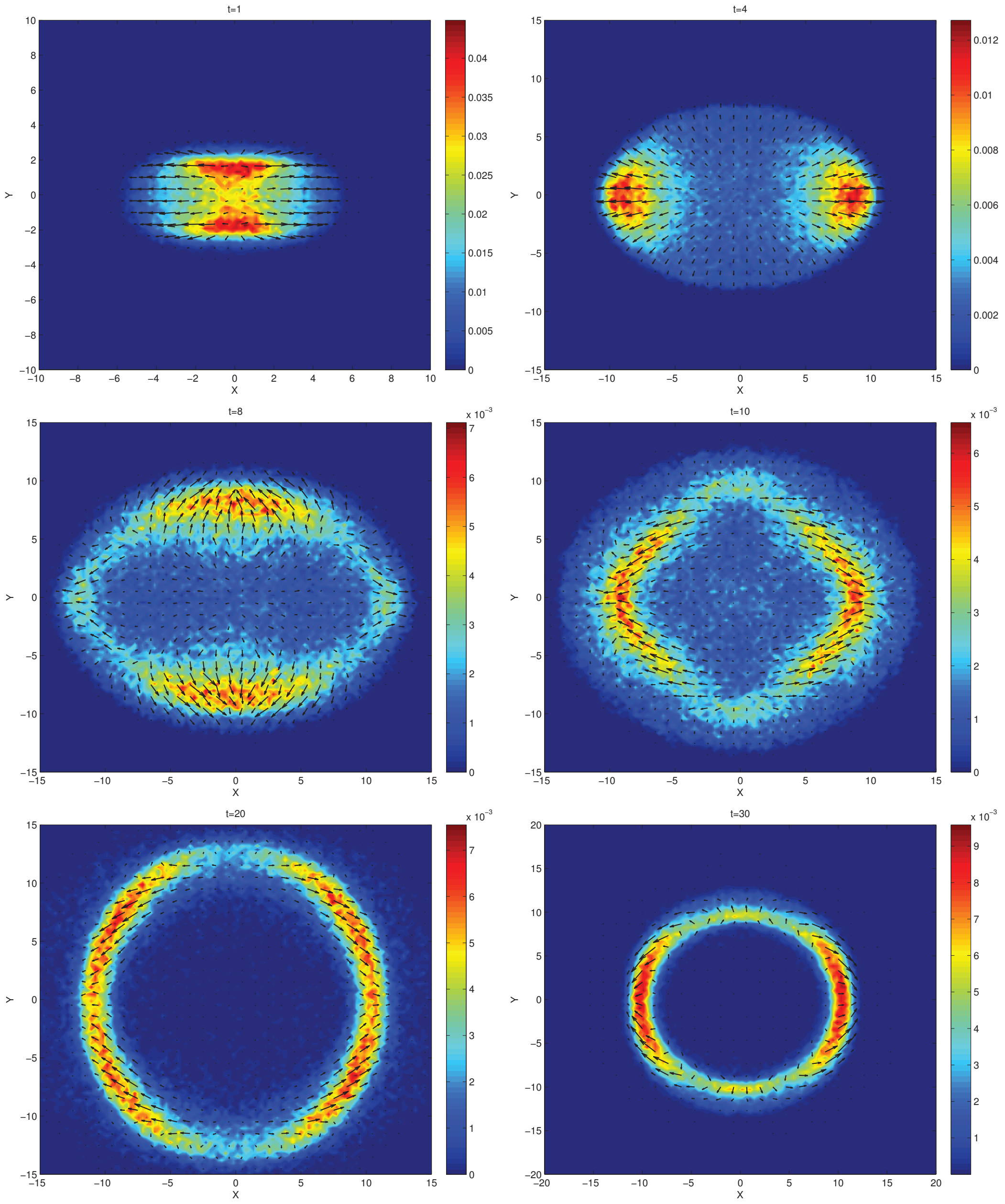


Figure 1.11: 2D Double Mill in D'Orsogna-Bertozzi et al. model at various times. Parameters in the attraction-repulsion potential are such that $C = C_R/C_A = 1.6$, $l = l_R/l_A = 0.025$, $\alpha = 0.7$, $\beta = 0.05$. Final configuration is reached after $t = 30$.

simulation we take

$$\phi(x) = \frac{d}{4}\left(\frac{|x|}{R_{roost}}\right)^4,$$

where R_{roost} gives the roosting area radius, and b is a constant weight. Other choice of this roost term are of course possible, we refer the interested reader to [51].

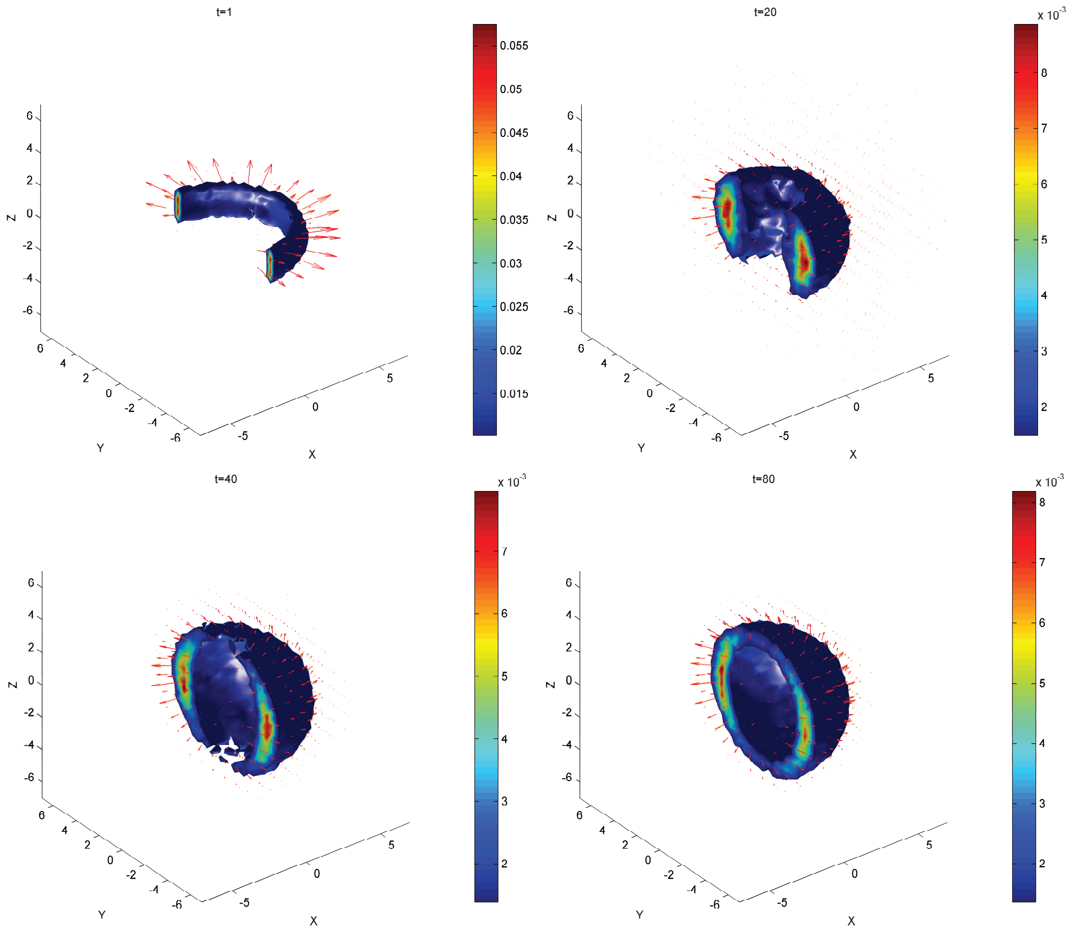


Figure 1.12: Evolution of the 3D mill in D'Orsogna-Bertozzi et al. model at different times.

Starting from the following initial data

$$f_0(x, y, z, v_x, v_y, v_z) = \frac{1}{2\pi\sigma^3} \exp\left\{\frac{1}{2\sigma^2}\left[(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2\right]\right\} \cdot$$

$$\cdot \frac{1}{\sqrt{2\pi\nu^2}} \exp\left\{\frac{1}{2\nu^2}\left[v_x^2 + v_y^2\right]\right\},$$

with $(x_0, y_0, z_0) = (-10, 10, 5)$, after a certain time the simulation shows a flock in stable equilibrium as an orbital motion around the roosting zone.

The simulation takes in account the following parameters $C = C_R/C_A = 30$, $l = l_R/l_A = 3/5$, $\alpha = 0.7$, $\beta = 0.05$ and the term of roosting force with parameters

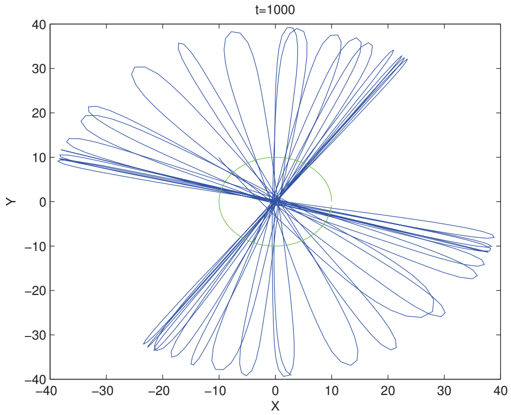


Figure 1.13: Trajectory of the center of mass in the roosting dynamic.

$R_{roost} = 10$ and $d = 1/10$. For a long time simulation the center of mass describe the trajectory depicted in Figure 1.5.4. Some configurations of the flock at different times obtained using $N = 200000$ individuals are reported in Figure 1.14.

1.6 Conclusions

Mathematical modeling of collective behavior involves the interaction of several individuals (of the order of millions) which may be computationally highly demanding. Here we focus on models for flocking and swarming where the particle interactions implies an $O(N^2)$ cost for N interacting objects. Using a probabilistic description based on a Boltzmann equation we show how it is possible to evaluate the interaction dynamic in only $O(N)$ computations. In particular we derive different approximation methods depending on a small parameter ε.

The building block of the method is given by classical binary collision simulations techniques for rarefied gas dynamic. Beside the presence of a further scaling parameter the resulting algorithms are fully meshless and can be applied to several different microscopic flocking/swarming models. Applications of the present ideas to other interacting particle systems and comparison with fast multipole methods are under study and will be presented elsewhere.

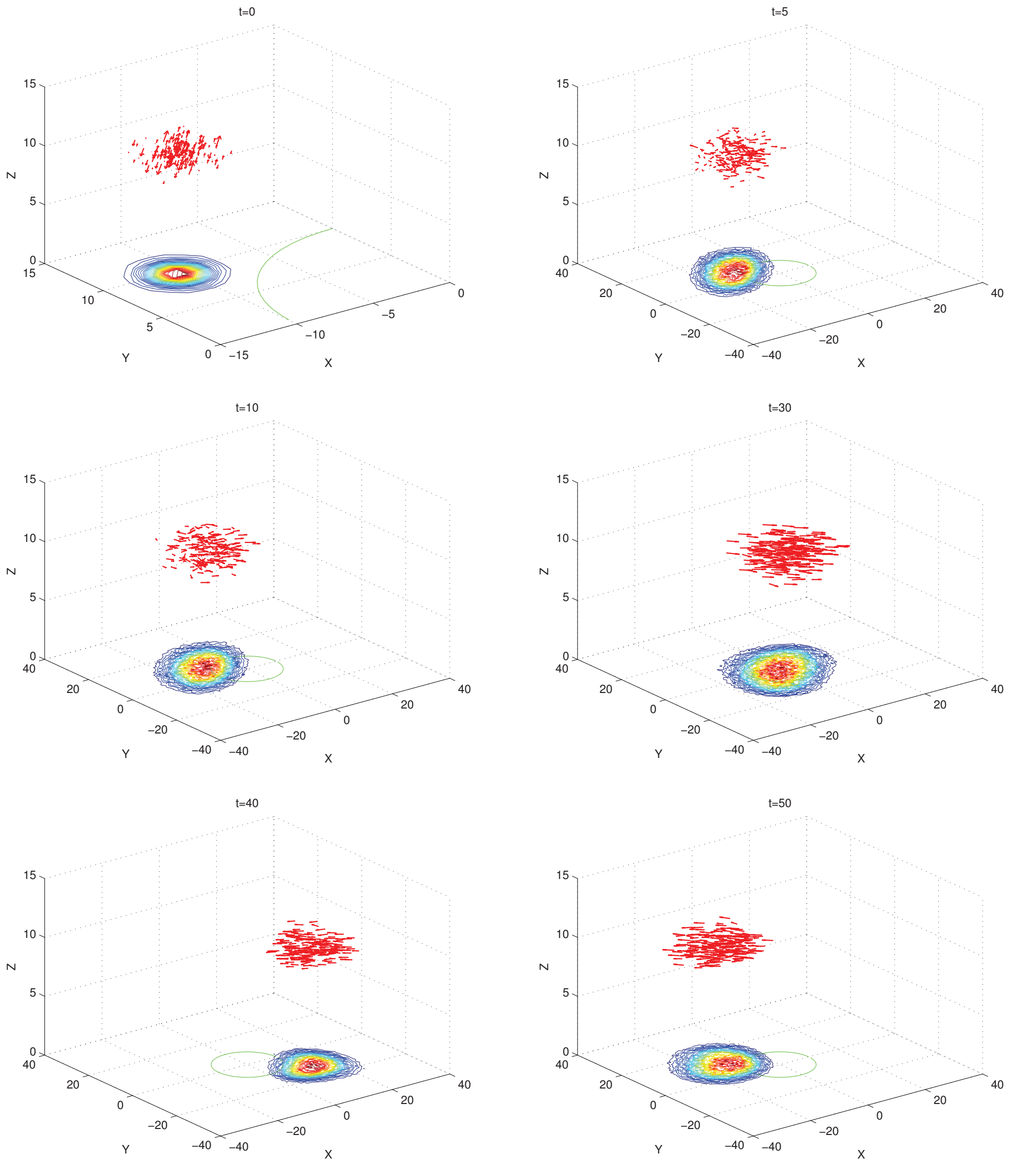


Figure 1.14: Evolution of the flock in 3D space, subject to a roosting force. The red arrow denotes the flock direction sampled from the initial population of $N = 200000$ individuals, the green circle represents the roosting area. We also add the density distribution of the whole flock projected over the plane (x, y).

CHAPTER 2

Stability analysis of flocking and mill rings for 2nd order models in swarming

2.1 Introduction

Individual-based models (IBMs) appear in biology, mathematics, physics, and engineering. They describe the motion of a collection of N individual entities, so the system is defined on a microscopic scale. IBMs are good models for some applications when the number of particles is reasonable. Nonetheless, when the number of particles is large it is more reasonable to use a continuum model. Some continuum models, like those described in [58, 50], are derived as a mean-field particle limit and lead to a mesoscopic, kinetic description of the problem. At this level, one looks at the probability density of finding particles at a certain position and velocity at a given time. Several related models have been proposed to describe the flocking of birds [45, 147, 19, 129], the formation of ant trails [82], the schooling of fish [104, 33, 21], swarms of bacteria [117], etc.

Each of these models include rules or mechanisms that describe the behavior of the individuals in the system. Such mechanisms can help to describe the influence of each individual on the others as a function of their relative position and velocity. The classical three zone model [10, 114] provides a well-known example. A three zone model describes the behavior of an individual in the following way: If two individuals are too close then they will prefer to have their own space (repulsion); When one individual is too far from the group it will prefer to socialize and therefore re-associate with the group (attraction); Finally, in the group, each individual tries to mimic the behavior of the other individuals (orientation). Other related models just consider rules for orientation, like the Vicsek model [158, 74]. In this case, there is a mechanism of self-propulsion in which each individual moves with constant speed and adopts the average direction among their local neighbors.

We focus our study on the analysis of two particular examples of IBMs. The first one is a self-propelled interacting particle model that was introduced in [127] and extensively studied in [80, 58]:

$$\begin{cases} \dot{x}_j = v_j \\ \dot{v}_j = S(|v_j|)v_j + \dfrac{1}{N}\displaystyle\sum_{\substack{l=1\\ l\neq j}}^{N} \nabla W(x_l - x_j) \end{cases}, \quad j = 1, \dots, N\,, \tag{2.1}$$

where W represents the social repulsive-attractive interaction potential assumed to be radial $W(x) = k(|x|)$. In our analysis we will consider the same *self-propulsion/ friction* term used in [80, 58],

$$S(|v_j|) = \alpha - \beta|v_j|^2, \qquad \alpha, \beta > 0.$$

Note that such a term gives a preferred asymptotic speed for the particles equal to $\sqrt{\alpha/\beta}$. In these references, the authors study (2.1) with interaction potential given by the so-called Morse potential

$$k(r) := C_A e^{-r/l_A} - C_R e^{-r/l_R},$$

with C_A, C_R denoting the attractive and repulsive strengths and l_A, l_R their respective length scales. The works [80, 58] find and describe several asymptotic behaviors for this system in 2D. They observed that flocking patterns and milling patterns can consist of particles distributed on a ring. They also observed that these patterns can occur when particles form into clusters instead of rings. In [47], a well-posedness theory is developed for (2.1) that proves the mean-field limit under smoothness assumptions on the potential. The authors show convergence of the particle model toward a measure solution of the corresponding kinetic equation.

We perform an analysis on the stability of flock rings and mill rings as asymptotic solutions for (2.1). A flock ring refers to a collection of individuals that lie equally distributed on a ring that translates with constant velocity, whereas a mill ring refers to a collection of individuals that lie equally distributed on a ring that rotates with constant angular velocity. The ring solution was recently studied in [119, 29] where the authors do a careful general linear stability analysis of the rings for the first order model

$$\dot{X}_j = \sum_{\substack{l=1\\ l\neq j}}^{N} \nabla W\left(X_l - X_j\right), \quad j = 1, \dots, N. \tag{2.2}$$

The analysis of (2.2) in [119, 29] is also used to study the the stability of mill rings in (2.1), whose existence was first demonstrated in [80]. Related pattern formation in the associated first order model has been studied in [161, 118].

Another second order model that we are going to study is

$$\begin{cases} \dot{x}_j = v_j \\ \dot{v}_j = \dfrac{1}{N}\displaystyle\sum_{l=1}^{N} H(x_j - x_l)(v_l - v_j) + \dfrac{1}{N}\displaystyle\sum_{\substack{l=1 \\ l\neq j}}^{N} \nabla W(x_l - x_j) \end{cases} \quad , \quad j = 1, \ldots, N \tag{2.3}$$

with $x_j, v_j \in \mathbb{R}^2$ where the velocity v_j is described by the Cucker-Smale alignment term, which quantifies the degree to which individuals align their velocities as a function of their relative positions. We perform our analysis in full generality for the parameter functions of the model H and W, but we will emphasize the results in some relevant cases. For instance, we consider the case of power law repulsive-attractive potentials [119, 18]

$$k(r) = \frac{r^a}{a} - \frac{r^b}{b}, \qquad a > b > 0. \tag{2.4}$$

For the Cucker-Smale alignment [67, 68, 100, 99, 53], a relevant case is $H(x) = g(|x|)$ with

$$g(r) = \frac{1}{(1+r^2)^{\gamma}}, \qquad \gamma > 0.$$

Note that flock solutions in the second order models (2.1) and (2.3) correspond to equilibria of the first order model (2.2). The main result of this work shows that flock solutions in the second order models (2.1) and (2.3) are *spectrally stable* if and only if the corresponding equilibrium is spectrally stable in the first order model (2.2). In other words, the linearized equations for (2.1) and (2.3) have an eigenvalue with positive real part if and only if the linearization for (2.2) has a positive eigenvalue. We therefore demonstrate a *spectral equivalence* between these three models.

To study the stability of the system of ODEs (2.1), we analyze the eigenvalues of the linearization of (2.1) at the equilibrium point in the comoving frame in its full generality. Unlike the first-order model (2.2), the zero solution to the linearized system associated to a flock solution of (2.1) is always unstable. This instability results from the fact that translational invariance implies the existence not only of an eigenvector with zero eigenvalue, but also an additional generalized eigenvector associated to the same zero eigenvalue of the linearized system (see Remark 2.3.1 for full details). We therefore identify all cases in which the linearized system has eigenvalues with zero real part as well as their corresponding generalized eigenspaces. This gives a complete characterization of stability at the linear level and of instability at the nonlinear level. This is shown in our main result that characterizes all eigenvalues with positive real part in the linearization of (2.1) in terms of the positive eigenvalues associated to the linearization of (2.2). The linearized analysis of the first order system (2.2) was already solved for the case of flock rings in [29] by using Fourier mode perturbations. Therefore, we can conclude with a full analysis of the instability of flock rings. However, an analysis of the stability of the family of flock solutions at the non-linear level is beyond the scope of this work. This analysis requires deep stability concepts from dynamical systems theory for invariant manifolds, and is under way in [56]. Finally, in the analysis of the linearization of (2.3) in the comoving frame, we use a similar strategy to [29] since the linearization using Fourier modes perturbations nicely decouples into a set of 4×4 ODE systems.

In addition to flock rings, other spatial shapes are possible as asymptotic solutions. When the flock ring is unstable we observe annular flocks, i.e. where the individuals' positions in the flock form an annulus, and we refer to this phenomenon as a *fattening instability*. We observe *clustering instabilities* as well, which occur when the individuals' positions in the flock highly concentrate in a small number of locations (usually lines or points). These patterns can be explained based on Theorem 2.3.1 due to the results in [17]. Moreover, these instabilities occur in the same way for both the first order and second order models, which provides a numerical demonstration of their spectral equivalence for flock rings.

Finally, we complement the analysis of the stability of asymptotic solutions for (2.1) by numerically studying mill rings. We extend the results of [29] to an exploration of mill configurations that appear with repulsive-attractive potentials. We numerically investigate the formation of fat mills, i.e. a group of individuals that fill an annular region while milling due to the repulsive force, and the formation of clusters by varying the asymptotic speed of the system. This instability induced by varying the asymptotic speed is quite interesting as it shows the rich pattern-forming structure for this model. In addition, we show some switching behaviors between flock and mill solutions that can occur as well.

The structure of the paper is as follows. In Section 2.2, we introduce the definitions of our main objects of study, the flock and mill rings. In Section 2.3, we do a linear stability analysis on the flock rings for models (2.1) and (2.3). We also explore the fattening and clustering instabilities. Finally, in Section 2.4 we analyze the instability with respect to the asymptotic speed for mill rings.

2.2 Ring Solutions

We begin by introducing the particular solutions of the particle model (2.1) and its continuum counterpart (see Figure 2.1) that we wish to study.

Definition 2.2.1 *We call a* flock ring*, the solution of* (2.1) *such that* $\{x_j\}_{j=1}^N$ *are equally distributed on a circle with a certain radius,* R *and* $\{v_j\}_{j=1}^N = u_0$*, with* $|u_0| = \sqrt{\alpha/\beta}$*.*

Definition 2.2.2 *We call a* mill ring, *the solution of* (2.1) *such that* $\{x_j\}_{j=1}^N$ *are equally distributed on a circle with a certain radius,* R *and* $\{v_j\}_{j=1}^N = u_j^0 = \sqrt{\frac{\alpha}{\beta}}\frac{x_j^\perp}{|x_j|}$ *with* $x_j^\perp$ *the orthogonal vector* $x_j^\perp = (x_{j,2}, -x_{j,1})$.

By abuse of notation, we will write $|u_0|$ for $|u_j^0|$ since $|u_j^0| = \sqrt{\alpha/\beta}$ for all $j = 1, \ldots, N$. Moreover, we will make use of notation $|u_0|$ for both flock and mill rings indistinctly.

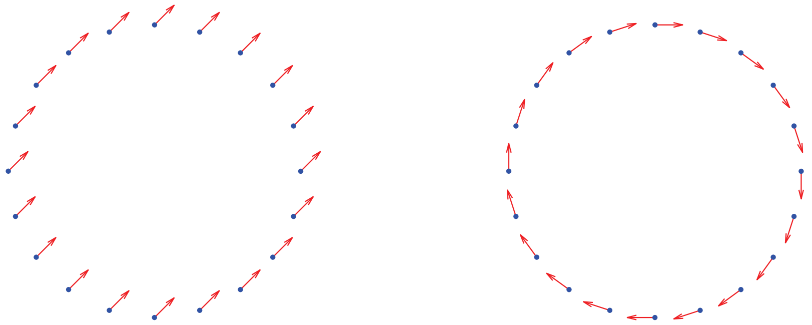


Figure 2.1: Flock and mill ring solutions.

2.2.1 Radius of Flock and Mill Ring Solutions

Throughout the paper we will identify $e^{i\theta} \equiv (\cos\theta, \sin\theta)$ and use x to identify either a two-dimensional vector or the corresponding complex number indistinctly when referring to ring solutions. In the case of mill rings, we are looking for a solution of the form

$$x_j(t) = R\left(\cos\left(\frac{2\pi}{N}j + \omega t\right), \sin\left(\frac{2\pi}{N}j + \omega t\right)\right) = Re^{i\frac{2\pi j}{N}}e^{i\omega t} \qquad (2.5)$$

The case of a flock ring in the comoving frame is equivalent to looking for a solution of the form (2.5) with $\omega = 0$. Plugging (2.5) into (2.1), for radial potentials $W(x) = k(|x|)$ we require that

$$\sum_{\substack{l=1\\l\neq j}}^N k'(|x_j - x_l|)\frac{x_j - x_l}{|x_j - x_l|} = 0. \qquad (2.6)$$

In the case of mill rings, we have

$$v_j(t) = \dot{x}_j(t) = R\omega i e^{i\theta_j}e^{i\omega t}, \quad \theta_j = \frac{2\pi j}{N},$$

and thus $R^2\omega^2 = \frac{\alpha}{\beta}$. Moreover, by taking the derivative we find

$$\dot{v}_j = -\omega^2 x_j = -\omega^2 \frac{1}{N}\sum_{\substack{l=1\\ l\neq j}}^{N}(x_l - x_j) = 0.$$

Plugging this into (2.1), we get

$$\sum_{\substack{l=1\\ l\neq j}}^{N}\left[k'(|x_l - x_j|)\frac{x_l - x_j}{|x_l - x_j|} - \omega^2(x_l - x_j)\right] = \sum_{\substack{l=1\\ l\neq j}}^{N}\tilde{k}'(|x_l - x_j|)\frac{x_l - x_j}{|x_l - x_j|} = 0,$$

with $\tilde{k}(r) = k(r) - \omega^2\frac{r^2}{2}$. Thus in order to find the radius for flock and mill rings, we need to solve equation (2.6) with potential $W(x) = \tilde{k}(|x|)$, and $\omega = 0$ or $\omega > 0$ respectively. This expression implies that the spatial shape has to balance attraction versus repulsion, as well as centrifugal forces in the case of mills. Now, a direct computation yields

$$|x_j - x_l| = 2R\sin\left(\frac{(l-j)\pi}{N}\right)$$

for all times. One can easily compute that

$$x_l - x_j = 2R\sin\left(\frac{p\pi}{N}\right)\begin{pmatrix} -\sin\left(\frac{p\pi}{N}\right) & \cos\left(\frac{p\pi}{N}\right) \\ \cos\left(\frac{p\pi}{N}\right) & \sin\left(\frac{p\pi}{N}\right)\end{pmatrix}\begin{pmatrix}\cos(\theta_j) \\ \sin(\theta_j)\end{pmatrix}, \quad p = l - j,$$

and

$$\sum_{\substack{l=1\\ l\neq j}}^{N}(x_j - x_l)\frac{\tilde{k}'(|x_j - x_l|)}{|x_j - x_l|} =$$

$$= \sum_{\substack{p=1-j\\ p\neq 0}}^{N-j}\begin{pmatrix} -\sin\left(\frac{p\pi}{N}\right)\cos(\theta_j) + \cos\left(\frac{p\pi}{N}\right)\sin(\theta_j) \\ \cos\left(\frac{p\pi}{N}\right)\cos(\theta_j) + \sin\left(\frac{p\pi}{N}\right)\sin(\theta_j)\end{pmatrix}\tilde{k}'\left(2R\sin\left(\frac{p\pi}{N}\right)\right)$$

$$= \begin{pmatrix}\cos(\theta_j) & \sin(\theta_j) \\ -\sin(\theta_j) & \cos(\theta_j)\end{pmatrix}\sum_{\substack{p=1-j\\ p\neq 0}}^{N-j}\begin{pmatrix} -\sin\left(\frac{p\pi}{N}\right) \\ \cos\left(\frac{p\pi}{N}\right)\end{pmatrix}\tilde{k}'\left(2R\sin\left(\frac{p\pi}{N}\right)\right).$$

Since the last sum is invariant by rotations, without loss of generality we fix $j = N$ and, by using periodicity, deduce that

$$\sum_{\substack{p=1-j\\ p\neq 0}}^{N-j}\begin{pmatrix} -\sin\left(\frac{p\pi}{N}\right) \\ \cos\left(\frac{p\pi}{N}\right)\end{pmatrix}\tilde{k}'\left(2R\sin\left(\frac{p\pi}{N}\right)\right) = \sum_{p=1}^{N-1}\begin{pmatrix} \sin\left(\frac{p\pi}{N}\right) \\ \cos\left(\frac{p\pi}{N}\right)\end{pmatrix}\tilde{k}'\left(2R\sin\left(\frac{p\pi}{N}\right)\right).$$

The sum of the cosine terms, evaluated on the uniform grid in $[0,\pi]$, vanishes, which leaves

$$\sum_{p=1}^{N-1}\begin{pmatrix}\sin\left(\frac{p\pi}{N}\right)\\ \\ \cos\left(\frac{p\pi}{N}\right)\end{pmatrix}\tilde{k}'\left(2R\sin\left(\frac{p\pi}{N}\right)\right)=\begin{pmatrix}\displaystyle\sum_{p=1}^{N-1}\sin\left(\frac{p\pi}{N}\right)\tilde{k}'\left(2R\sin\left(\frac{p\pi}{N}\right)\right)\\ \\ 0\end{pmatrix}. \tag{2.7}$$

As a conclusion, the radius of a flock or mill ring solution is characterized by

$$\sum_{p=1}^{N-1}\sin\left(\frac{p\pi}{N}\right)\tilde{k}'\left(2R\sin\left(\frac{p\pi}{N}\right)\right)=0.$$

For general potentials there can be more than one radius R for a flock or mill solution. In the case of the power law potentials (2.4), we will argue below that there is only one solution. Condition (2.7) reads

$$(2R)^{a-1}\frac{1}{N}\sum_{p=0}^{N-1}\sin^a\left(\frac{p\pi}{N}\right)-(2R)^{b-1}\frac{1}{N}\sum_{p=0}^{N-1}\sin^b\left(\frac{p\pi}{N}\right)-2R\omega^2\frac{1}{N}\sum_{p=0}^{N-1}\sin^2\left(\frac{p\pi}{N}\right)=0. \tag{2.8}$$

To prove uniqueness, we notice that the function $f(r)=C_1r^a-C_2r^b-C_3$ with $a>b>0$ and $C_1>C_2>0$, $C_3>0$, has only one zero. Computing the first derivative and looking for critical points, we obtain $r_1=0$ and $r_2^{a-b}=\frac{C_2b}{C_1a}$. Taking the second derivative and evaluating at r_2, one obtains $f''(r_2)=r_2^{b-2}C_2b(a-b)>0$, so r_2 is a local minimum. For $0<r<r_2$, one has $f'(r)<0$ whereas for all $r\in(r_2,+\infty)$, one has $f'(r)>0$. We may therefore conclude that $f(r)$ has a unique zero. Notice that the solution to (2.8) depends on the number of particles, so we will use the notation $R=R(N)$ to indicate this dependency.

2.2.2 Radius of the Flock and Mill Ring Solutions as $N\to\infty$

In this subsection we characterize the radius of flock and mill rings for continuum models arising as $N\to\infty$ for the power-law case. We introduce the function

$$\psi_\alpha(s)=\frac{1}{\pi}\int_0^\pi(1-s\cos\theta)(1+s^2-2s\cos\theta)^{\frac{\alpha-2}{2}}\,d\theta,$$

already analyzed in [18]. A change of variables in the previous function shows that

$$\psi_\alpha(1)=\frac{2^{\alpha-1}}{\pi}B\left(\frac{\alpha+1}{2},\frac{1}{2}\right), \tag{2.9}$$

where $B(\cdot,\cdot)$ stands for the Beta function defined as

$$B(x,y) = 2\int_0^{\pi/2} (\cos t)^{2x-1}(\sin t)^{2y-1}\,dt,$$

with $x, y > 0$.

Lemma 2.2.1 *If $N \to \infty$ then $R(N) \to R_{ab}(|u_0|)$ where $R_{ab}(|u_0|)$ is the solution of the following equation:*

$$\psi_a(1)R^{a-1} - \psi_b(1)R^{b-1} - \omega^2 R = 0.$$

Proof 2.2.1 *Let us start by studying the term $2^{a-1}\frac{1}{N}\sum_{p=0}^{N-1}\sin^a(\frac{p\pi}{N})$. Multiplying and dividing by π we obtain the following equality*

$$\lim_{N\to\infty} 2^{a-1}\frac{1}{\pi}\left(\frac{\pi}{N}\sum_{p=0}^{N-1}\sin^a\left(\frac{p\pi}{N}\right)\right) = 2^{a-1}\frac{1}{\pi}\int_0^{\pi}\sin^a(x)\,dx = 2^a\frac{1}{\pi}\int_0^{\pi/2}\sin^a(x)\,dx. \tag{2.10}$$

Now, we use the expression for the Beta function described above with $x = \frac{1}{2}$, $y = \frac{a+1}{2}$, and using that $B(x,y) = B(y,x)$ in (2.10) together with (2.9) to obtain

$$\lim_{N\to\infty} 2^{a-1}\frac{1}{\pi}\left(\frac{\pi}{N}\sum_{p=0}^{N-1}\sin^a\left(\frac{p\pi}{N}\right)\right) = 2^{a-1}\frac{1}{\pi}B\left(\frac{a+1}{2},\frac{1}{2}\right) = \psi_a(1).$$

The same reasoning works by changing a for b in the second term in (2.8). For the third term we use the fact that we can compute the exact sum

$$\frac{2}{N}\sum_{p=0}^{N-1}\sin^2\left(\frac{p\pi}{N}\right) = 1.$$

Remark 2.2.1 *In the case of flock rings $\omega = 0$, their radius is determined by the radius of the aggregation ring found in [18]*

$$R(N) \to R_{ab} = \frac{1}{2}\left(\frac{B(\frac{b+1}{2},\frac{1}{2})}{B(\frac{a+1}{2},\frac{1}{2})}\right)^{\frac{1}{a-b}} \quad \textit{as} \quad N \to \infty.$$

Remark 2.2.2 *Let $W(x) = k(|x|)$ be a general interaction potential. Call $Q(r) = -k'(r)/r$. Then the radius of the ring is determined by*

$$\int_0^{\frac{\pi}{2}} Q(2R\sin(\theta))\sin^2(\theta)\,d\theta = 0,$$

as shown in [29].

Remark 2.2.3 *The corresponding continuum model to the particle system* (2.1)*, as proven in* [47]*, is given by the kinetic equation*

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + \mathrm{div}_v[(\alpha - \beta|v|^2)vf)] - \mathrm{div}_v[(\nabla_x W * \rho)f] = 0, \tag{2.11}$$

where

$$\rho(t,x) = \int_{\mathbb{R}^2} f(t,x,v)\,dv.$$

Here, f represents the probability of finding individuals at a time t at the point (x,v) and ρ is the spatial density of individuals. It was shown in [50] *that singular solutions of the type*

$$f(t,x,v) = \rho(t,x)\delta\left(v - u_0\right), \qquad f(t,x,v) = \rho(t,x)\delta\left(v - \sqrt{\frac{\alpha}{\beta}}\frac{x^\perp}{x}\right),$$

with $\rho(t,x)$ the uniform distribution on a ring, are weak solutions of the kinetic model (2.11)*, called the flock and mill ring continuous solutions respectively.*

2.3 Linear Stability Analysis for Flock Solutions

We will now focus on the stability analysis of flock rings in full generality. Later on, we will leverage these results together with the careful stability analysis of ring solutions of the aggregation equation, performed in [29], to study their stability in terms of the parameters (a, b, u_0) of the model for the particular case of power-law potentials.

2.3.1 Stability of Flock Solutions to (2.1)

In order to address stability of flock solutions to (2.1), we must first perform a change of variables to the comoving frame

$$\begin{cases} y_j(t) = x_j(t) - u_0 t \\ z_j(t) = v_j(t) - u_0 \end{cases} \qquad j = 1, \dots, N, \tag{2.12}$$

where u_0 is the asymptotic velocity of a fixed flock ring. Under this change of variables, the system therefore (2.1) reads

$$\begin{cases} \dfrac{d}{dt} y_j = v_j - u_0 = z_j \\ \dfrac{d}{dt} z_j = \underbrace{(\alpha - \beta|z_j + u_0|^2)}_{S_0(z_j)}(z_j + u_0) + \dfrac{1}{N}\displaystyle\sum_{\substack{l=1 \\ l\neq j}}^{N} k'(|y_l - y_j|)\frac{y_l - y_j}{|y_l - y_j|} \end{cases}, \quad j = 1, \dots, N.$$

A flock ring can then be characterized as a stationary solution of this system that takes the form $\left(y_j^0, z_j^0\right) = \left(Re^{i\theta_j}, 0\right)$, where $\theta_j = \frac{2\pi j}{N}$ for $j = 1, \ldots, N$. This stationary solution satisfies

$$S_0(z_j^0) = 0, \quad \sum_{\substack{l=1 \\ l \neq j}}^{N} k'(|y_j^0 - y_l^0|) \frac{(y_l^0 - y_j^0)}{|y_l^0 - y_j^0|} = 0.$$

We may now proceed to linearize this equation around a flock solution. We consider arbitrarily perturbed solutions of the form $y_j = y_j^0 + h_j(t)$ for $h_j \in \mathbb{R}^2$ with $|h_j| \ll 1$, which to leading order yields

$$h_j'' = -2\beta u_0 u_0^T h_j' + \frac{1}{N} \sum_{\substack{l=1 \\ l \neq j}}^{N} M(y_i^0, y_j^0)(h_l - h_j)$$

where M is the matrix defined as follows

$$M(y_i^0, y_j^0) = \frac{k'(|y_{jl}^0|)}{|y_{jl}^0|} \left(\mathrm{Id} - \frac{y_{jl}^0}{|y_{jl}^0|} \frac{(y_{jl}^0)^T}{|y_{jl}^0|} \right) + k''(|y_{jl}^0|) \frac{y_{jl}^0}{|y_{jl}^0|} \frac{(y_{jl}^0)^T}{|y_{jl}^0|},$$

and $y_{jl}^0 := y_j^0 - y_l^0$. Note that the final term on the right hand side coincides with the linearization of the first order model (2.2) around the equilibrium $\{y_j^0\}_{j=1}^N$ that defines the flock. If we put $\mathbf{h} := (h_1, \ldots, h_N)^T \in \mathbb{R}^{2N}$ and $\mathbf{h}' := (h_1', \ldots, h_N')^T \in \mathbb{R}^{2N}$ then we may write the preceeding linear system in compact form as

$$\frac{d}{dt} \begin{pmatrix} \mathbf{h} \\ \mathbf{h}' \end{pmatrix} = \begin{pmatrix} 0 & \mathrm{Id} \\ & \\ \mathbf{M} & -2\beta U \end{pmatrix} \begin{pmatrix} \mathbf{h} \\ \mathbf{h}' \end{pmatrix}.$$

An arbitrary perturbation for general flocks therefore leads to a Jacobian matrix of the form

$$L := \begin{pmatrix} 0 & \mathrm{Id} \\ & \\ \mathbf{M} & -2\beta U \end{pmatrix},$$

where the partition into $2N \times 2N$ sub-blocks reflects the distinction between position and velocity contributions to the Jacobian: The symmetric matrix $\mathbf{M}$ is the $2N \times 2N$ Hessian that results from linearizing the first order system (2.2) about a given flock configuration, whereas U denotes a block-diagonal matrix with N blocks of the 2×2 matrix $u_0 u_0^T$ along the diagonal. By rotational invariance we can reduce to the case $u_0 = e_1 = (1, 0)$, so that the block matrix U acts on $\mathbf{x} = (x_1, \ldots, x_N)^T \in \mathbb{R}^{2N}$, $x_i \in \mathbb{R}^2$, according to the relation

$$(U\mathbf{x})_i = \begin{pmatrix} \langle x_i, e_1 \rangle \\ 0 \end{pmatrix}.$$

We now turn to the task of characterizing the eigenvalues of L in terms of the eigenvalues of $\mathbf{M}$. In other words, we aim to characterize the stability of a flock in terms of the stability of its spatial shape as a solution to the first order model. To fix the notation, we write the eigenvalue problem for the flock as

$$\lambda \begin{pmatrix} \mathbf{x} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} 0 & \mathrm{Id} \\ \mathbf{M} & -2\beta U \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{v} \end{pmatrix} = L \begin{pmatrix} \mathbf{x} \\ \mathbf{v} \end{pmatrix}, \tag{2.13}$$

where the matrix $\mathbf{M}$ determines the stability of the flocking configuration as a solution of the first order model. For any given eigenvector $(\mathbf{x}, \mathbf{v}) \in \mathbb{C}^{2N} \times \mathbb{C}^{2N}$ of the full system (2.13), we always assume the normalization $\mathbf{x}^*\mathbf{x} = 1$. Substituting the first equation $\lambda\mathbf{x} = \mathbf{v}$ into the second equation yields the equivalent statement

$$\lambda^2\mathbf{x} + 2\beta\lambda U\mathbf{x} - \mathbf{M}\mathbf{x} = 0. \tag{2.14}$$

Let $|\mathbf{x}|_2$ denote the semi-norm on $\mathbb{C}^{2N}$ defined according to

$$|\mathbf{x}|_2^2 := \sum_{i=1}^{N} |\langle x_i, e_1\rangle|^2,$$

and let $E^N \cong \mathbb{C}^N$ denote the subspace

$$E^N := \left\{\mathbf{x} \in \mathbb{C}^{2N} : |\mathbf{x}|_2 = 0\right\} = \ker(U).$$

Premultiplying by $\mathbf{x}^*$, the fact that $\mathbf{x}^*U\mathbf{x} = |\mathbf{x}|_2^2$, the normalization on $\mathbf{x}$ and the quadratic formula combine to imply the key identity

$$\lambda = -\beta|\mathbf{x}|_2^2 \pm \sqrt{\beta^2|\mathbf{x}|_2^4 + \mathbf{x}^*\mathbf{M}\mathbf{x}}. \tag{2.15}$$

As $\mathbf{M}$ is symmetric, we may write its $2N$ real eigenvalues and corresponding normalized ($\mathbf{x}^*\mathbf{x} = 1$) eigenvectors as

$$\mu_{2N} \leqslant \mu_{2N-1} \leqslant \cdots \leqslant \mu_2 \leqslant \mu_1 \qquad \mathbf{M}\mathbf{x}_i = \mu_i\mathbf{x}_i.$$

The notation $a_L(\lambda)$, $a_{\mathbf{M}}(\mu)$ will denote the algebraic multiplicities of λ, μ as eigenvalues of their respective matrices. The bulk of the analysis lies in characterizing the eigenvalues λ of the full system (2.13) that have $\Re(\lambda) = 0$.

Lemma 2.3.1 *Let λ denote an eigenvalue of (2.13). Then $\Re(\lambda) = 0$ and $\Im(\lambda) \neq 0$ if and only if $\lambda = \pm i\sqrt{-\mu_l}$ for some l with $\mu_l < 0$ and $\mathbf{x}_l \in E^N$. The eigenspace consists only of eigenvectors.*

Proof 2.3.1 *If $\mathbf{x}_l \in E^N$ then (2.14) reads $\lambda^2 \mathbf{x}_l = \mathbf{M}\mathbf{x}_l$, or equivalently $\lambda^2 = \mu_l$. To have $\Im(\lambda) \neq 0$ then requires $\mu_l < 0$. Conversely, if $(\mathbf{x}, \lambda\mathbf{x})$ denotes an eigenvector with $\Re(\lambda) = 0$ and $\Im(\lambda) \neq 0$, the formula (2.15) implies that necessarily $\mathbf{x} \in E^N$, and therefore $\mathbf{M}\mathbf{x} = \lambda^2\mathbf{x}$. Thus $\lambda^2 = \mu_l$ for some $\mu_l < 0$.*

To show the last statement, suppose a generalized eigenvector existed that is not an eigenvector. Then there exists an eigenvector $(\mathbf{x}, \lambda\mathbf{x})$ with $\mathbf{x} \in E^N$ so that the system of equations

$$\begin{pmatrix} -\lambda\mathrm{Id} & \mathrm{Id} \\ \mathbf{M} & -2\beta U - \lambda\mathrm{Id} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{x} \\ \lambda\mathbf{x} \end{pmatrix} \tag{2.16}$$

has a non-trivial solution. Substituting the first equation $\mathbf{w} = \lambda\mathbf{u} + \mathbf{x}$ into the second equation, then pre-multiplying by $\mathbf{x}^*$ demonstrates

$$\begin{aligned} \mathbf{M}\mathbf{u} - 2\beta U\mathbf{w} &= 2\lambda\mathbf{x} + \lambda^2\mathbf{u} \\ \mathbf{x}^*\mathbf{M}\mathbf{u} &= 2\lambda + \lambda^2\mathbf{x}^*\mathbf{u}. \end{aligned}$$

The last line follows as $\mathbf{x}^*\mathbf{x} = 1$ and $\mathbf{x} \in E^N = \ker(U)$. The symmetry of $\mathbf{M}$ and the fact that $\mathbf{M}\mathbf{x} = \lambda^2\mathbf{x}$ combine to show $\mathbf{x}^*\mathbf{M}\mathbf{u} = \lambda^2\mathbf{x}^*\mathbf{u}$. Thus $\lambda = 0$, leading to a contradiction.

Lemma 2.3.2 *Let $\beta > 0$. Then $\lambda = 0$ is an eigenvalue of (2.13) and $(\mathbf{x}, \mathbf{0})$ is a corresponding eigenvector if and only if $\mathbf{M}\mathbf{x} = \mathbf{0}$. If $\mathbf{x} \in E^N$ then $(\mathbf{x}, \mathbf{0})$ generates a single generalized eigenvector, whereas if $\mathbf{x} \notin E^N$ then $(\mathbf{x}, \mathbf{0})$ generates no generalized eigenvectors.*

Proof 2.3.2 *The first statement follows trivially from (2.14). To see the second statement, consider the system of equations (2.16) with $\lambda = 0$. This reduces to the equations $\mathbf{w} = \mathbf{x}$ and*

$$\mathbf{M}\mathbf{u} = 2\beta U\mathbf{x},$$

which by premultiplying by $\mathbf{x}^*$ as before and using the fact that $\mathbf{M}\mathbf{x} = \mathbf{0}$ necessitates $\mathbf{x} \in E^N$ as $\beta > 0$. If indeed $\mathbf{x} \in E^N$ then any $\mathbf{u} \in \ker(\mathbf{M})$ suffices. Without loss of generality, take $\mathbf{u} = \mathbf{x}$ itself. If $(\mathbf{x}, \mathbf{0})$ generates a second generalized eigenvector then the system of equations

$$\begin{pmatrix} 0 & \mathrm{Id} \\ \mathbf{M} & -2\beta U \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{x} \\ \mathbf{x} \end{pmatrix}$$

has a non-trivial solution. As then $\mathbf{w} = \mathbf{x}$ and $\mathbf{x} \in E^N$ this reads $\mathbf{M}\mathbf{u} = \mathbf{x}$. Premultiplying one last time by $\mathbf{x}$, the facts that $\mathbf{M}\mathbf{x} = \mathbf{0}$ and $\mathbf{x}^*\mathbf{x} = 1$ combine to produce the contradiction $0 = 1$.

This lemma yields, as a corollary, the algebraic multiplicity $a_L(0)$ of zero as an eigenvalue of the second order system.

Corollary 2.3.1 *Let $\beta > 0$. Then*

$$a_L(0) = \dim(\ker(\mathbf{M}) \cap E^N) + \dim(\ker(\mathbf{M})).$$

Let $a_{\mathbf{M},\perp}(0) := \dim(\ker(\mathbf{M}) \cap E^N)$, so that $a_L(0) = a_{\mathbf{M},\perp}(0) + a_{\mathbf{M}}(0)$. Note that neither quantity depends on β, and the conclusion holds whenever β is positive. Thus, if $\beta \in (0, \infty)$ it follows that $a_L(0)$ is constant. Moreover, Lemma 2.3.1 holds uniformly in β as well. Let $i_1 < i_2 < \cdots < i_l \leqslant 2N$ denote those (possibly non-existent) indices where $\mu_{i_j} < 0$ has an eigenvector $\mathbf{x}_{i_j} \in E^N$. The two lemmas then combine to show:

Corollary 2.3.2 *Let $\beta > 0$. Then*

$$\det(L - \lambda \mathrm{Id}) = \lambda^{a_{\mathbf{M},\perp}(0)+a_{\mathbf{M}}(0)} \Pi_{j=1}^{l}(\lambda^2 - \mu_{i_j}) p_\beta(\lambda).$$

The roots of the polynomial $p_\beta(\lambda)$ all have non-zero real part.

This corollary, along with the formula (2.15), suffice to establish the desired result:

Theorem 2.3.1 *(Spectral Equivalence) The linearized second order system around the flock ring solution (2.1) has an eigenvalue with positive real part if and only if the linearized first order system around the ring solution has a positive eigenvalue.*

Proof 2.3.3 *Suppose first that $\mu_1 \leqslant 0$. Then $\mathbf{x}^*\mathbf{M}\mathbf{x} \leqslant 0$ for any $\mathbf{x}$, whence all eigenvalues λ of L have non-positive real part due to (2.15). Conversely, suppose $\mu_1 > 0$ and let $\mathcal{A}$ denote the set*

$$\mathcal{A} := \left\{ \beta \in [0, \infty) : \max_{\lambda \in \sigma(L)} \Re(\lambda) > 0 \right\}.$$

Note that $0 \in \mathcal{A}$ due to (2.15). Indeed, then $(\mathbf{x}_1, \sqrt{\mu_1}\mathbf{x}_1)$ defines an eigenvector with eigenvalue $\lambda = \sqrt{\mu_1} > 0$. By continuous dependence of the eigenvalues of L on β, it follows that $\mathcal{A}$ is relatively open. To show that it is also relatively closed, let $\beta_l \in \mathcal{A}$ and $\beta_l \to \beta_0 \in (0, \infty)$. Up to extraction of subsequences, it follows that there exists a corresponding sequence λ_l of eigenvalues with $\Re(\lambda_l) > 0$ converging to some λ_0 with $\Re(\lambda_0) \geqslant 0$. Moreover, by continuous dependence of the coefficients of $p_\beta(\lambda)$ on β, the roots of $p_{\beta_l}(\lambda)$ converge to roots of $p_{\beta_0}(\lambda)$. Thus $p_{\beta_0}(\lambda_0) = 0$. As no such root can have zero real part by corollary 2.3.2, $\Re(\lambda_0) > 0$ and $\beta_0 \in \mathcal{A}$. As $\mathcal{A} \neq \varnothing$ it follows that $\mathcal{A} = [0, \infty)$ as desired.

Remark 2.3.1 *As an artifact of translation invariance in the first order model, the vector defined by* $\mathbf{e}_2 := (0, 1, \dots, 0, 1)^T \in \mathbb{R}^{2N}$ *always defines an eigenvector of* $\mathbf{M}$ *with eigenvalue zero. Due to the fact that* $\mathbf{e}_2 \in E^N$*, Lemma 2.3.2 implies that* $(\mathbf{e}_2, \mathbf{e}_2)$ *furnishes a generalized eigenvector with eigenvalue zero. Thus the linear system (2.13) that results from linearization of (2.1) about a flock always has an unstable zero solution. This type of instability does not occur for the first order model and is therefore unique to the second order case.*

We may now specify this result to the particular case of flock rings. Theorem 2.3.1 implies that spectral stability of a flock ring is equivalent to spectral stability of a ring solution $y_j = Re^{i\theta_j}$, $\theta_j = \frac{2\pi i j}{N}$ to the first order model (2.2). Moreover, the analysis in [29] shows that the stability analysis of ring solutions to (2.2) reduces to a study of the decoupled set of 2×2 eigenvalue problems

$$\lambda \begin{pmatrix} \xi_+ \\ \xi_- \end{pmatrix} = \underbrace{\begin{pmatrix} I_1(m) & I_2(m) \\ I_2(m) & I_1(-m) \end{pmatrix}}_{M} \begin{pmatrix} \xi_+ \\ \xi_- \end{pmatrix} \qquad 1 \leqslant m \leqslant N. \tag{2.17}$$

Here the matrix entries $I_1(m)$ and $I_2(m)$ are defined by

$$I_1(m) := 4 \sum_{p=1}^{N/2} G_1\left(\frac{\pi p}{N}\right) \sin^2\left(\frac{(m+1)\pi p}{N}\right) \tag{2.18}$$

$$I_2(m) := 4 \sum_{p=1}^{N/2} G_2\left(\frac{\pi p}{N}\right) \left[\sin^2\left(\frac{\pi p}{N}\right) - \sin^2\left(\frac{m\pi p}{N}\right)\right], \tag{2.19}$$

and for power-law potentials $k(r) = r^a/a - r^b/b$ the functions $G_i(\phi)$ are given by

$$G_1(\phi) := \frac{1}{2N}\left[-a(2R|\sin\phi|)^{a-2} + b(2R|\sin\phi|)^{b-2}\right],$$
$$G_2(\phi) := \frac{1}{2N}\left[-(a-2)(2R|\sin\phi|)^{a-2} + (b-2)(2R|\sin\phi|)^{b-2}\right].$$

This result follows by considering m-mode perturbations to the ring equilibrium, i.e. perturbations of the form $Re^{i\theta_j}(1 + h_j)$ for

$$h_j = \xi_+ e^{im\theta_j} + \xi_- e^{-im\theta_j},$$

so that a study of stability decouples into a study of individual Fourier modes. We may therefore conclude the following corollary.

Corollary 2.3.3 *A flock ring to (2.1) is spectrally stable if and only if the ring solution to the first order model (2.2) is spectrally stable with respect to all* m*-mode perturbations.*

2.3.2 Numerical Tests

In this section we perform some numerical computations to show stability regions for the flock ring. Moreover, we enrich the previous analysis by showing that the same type of instabilities occur in both the second order model and the first order model. Specifically, when parameters are chosen outside the stability regions either clustering or fattening instabilities occur simultaneously in both models, as one can formally argue due to their spectral equivalence proved in Corollary 2.3.3.

Due to Theorem 2.3.1 and Corollary 2.3.3 we are reduced to a study of the trace and determinant of the matrix M in (2.17) to characterize flock stability. Note that for fixed values of N and m the determinant of M is a function of the parameters a and b, so that we may write

$$\begin{aligned} D(a,b) :=& \det(M) = I_1(m)I_1(-m) - (I_2(m))^2, \\ T(a,b) :=& \text{trace}(M) = I_1(m) + I_1(-m). \end{aligned}$$

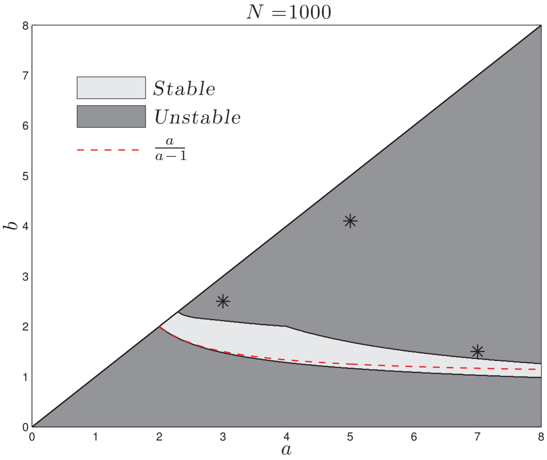


Figure 2.2: Stability areas for flock ring solutions for $N = 1000$. Markers ($\ast$) indicate the explored parameters in Table 2.1 .

Remark 2.3.2 *Using the results of* [29, Theorem 3.1] *one is able to estimate the asymptotic value of the determinant of M. In our case, using $W(x) = \frac{|x|^a}{a} - \frac{|x|^b}{b}$ one obtains that*

$$\det(M) \sim Cm^{-b+1} \quad as \quad m \to \infty,$$

where $C > 0$ and $b \in (1,2) \cup (4,6) \cup (8,10) \cup \cdots$. In these cases $\det(M) > 0$ *and* $\text{trace}(M) < 0$. *Moreover, this result shows that there is no spectral gap since* $\det(M) \to 0$ *as $m \to \infty$.*

In Figure 2.2, we compute the stability area as a function of a and b for $N = 1000$. To do so, we compute the intersection of all stability areas for $m \geqslant 2$. It can be observed from our tests that the stability area shrinks when the number of particles increases. Moreover, it is observed that in the limit when $N \to \infty$, the lower boundary of the stability region converges to the dashed line. The red dashed line is the curve $b = \frac{a}{a-1}$ that corresponds to the $m = +\infty$ mode. This curve is the separatrix of the ins/stability regions for the continuous delta ring of the first order continuum model, studied in [119, 18].

Cluster Formation

The formation of clusters occurs when the repulsion strength is small. In other words, this phenomenon depends on how singular the potential is at the origin. We show the bifurcation diagram for the phase transition between equally distributed flocks and flocks that exhibit cluster formation. Figure 2.3 is obtained by first using

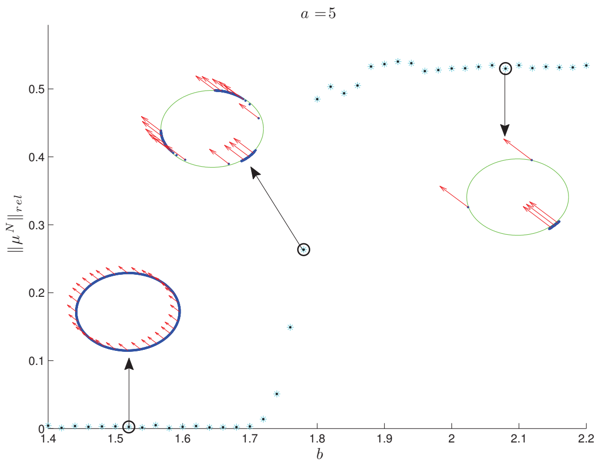


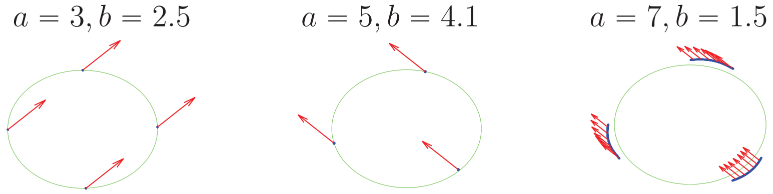
Figure 2.3: Bifurcation diagram for cluster formation at $T_f = 500$, with $N = 1000$ particles, $a = 5$, $|u_0| = 2.5$.

$N = 1000$ particles equally distributed on the stable circle with all their velocities aligned. We let them evolve until $T_f = 500$. We fix the parameters $|u_0| = 2.5$, $a = 5$ and vary b along the axis. The vertical axis represents the increment of the following norm

$$\|\mu^N\|_{rel} = \frac{\|\mu^N - \mu_0^N\|_2}{\|\mu_0^N\|_2}$$

with increasing b, where μ_0^N is the uniform distribution of N particles along the stable ring and μ^N the distribution at time T_f. We therefore measure the relative distance of a computed flock from the flock ring. Simulations are performed with *MATLAB* and the evolution of the system of ODEs is solved with the *ode45* routine with adaptive time step. Table 2.1 illustrates different final states depending on a, b. The parameters choice depicts some of the patterns observed for the first order model in [17].

Table 2.1: Long time simulations with $N = 1000$ particles. The location of parameter values are marked in Figure 2.2.



Fattening Formation

We show the transition diagram between a flock on a ring and a flock on an annulus. In this case, the fattening phenomenon occurs when the parameters of the potential cross the lower boundary of the stability region. We numerically characterize this behavior in a similar way as done in the previous subsection.

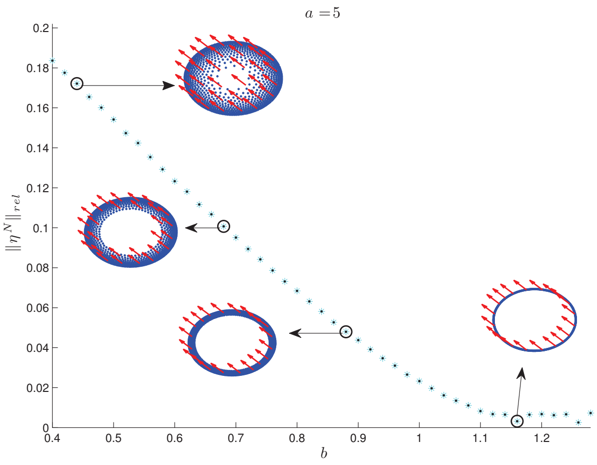


Figure 2.4: Bifurcation diagram for fattening instability at $T_f = 500$ with N=1000 particles, $a = 5$, $|u_0| = 2.5$.

Figure 2.4 is obtained by first starting with $N = 1000$ particles equally distributed on the stable ring already in the steady state with all the velocities aligned.

We then let them evolve until $T_f = 500$ as before. We fix the parameters $|u_0| = 2.5$, $a = 5$ and vary b along the axis. The vertical axis represents the increment of the relative distance from the mill ring solution with respect to the following norm

$$\|\eta^N\|_{rel} = \frac{\|\eta^N - \eta_0^N\|_2}{\|\eta_0^N\|_2},$$

where η_0^N represents the average distance from the center of mass for N particles in a flock ring formation, i.e., $\eta_0^N = R$ and η^N is the average distance from the center of mass at time T_f.

2.3.3 Stability of Flock Solutions to (2.3)

As in Section 2.3.1, we perform a stability analysis of flock solutions for the model (2.3). If we use the same change of variables as in (2.12), then system (2.3) reads

$$\begin{cases} \dot{y}_j = v_j - u_0 = z_j \\ \dot{z}_j = \dfrac{1}{N}\displaystyle\sum_{l=1}^{N} H(|y_l - y_j|)(z_l - z_j) + \dfrac{1}{N}\sum_{\substack{l=1\\ l\neq j}}^{N} \nabla W(y_l - y_j), \end{cases} \qquad j = 1, \dots, N. \tag{2.20}$$

From now on, we will identify vectors in the plane with complex numbers to perform the linearization of this system for m-mode perturbations of the flock ring. Unlike in the previous model (2.1), these particular perturbations allow us to conveniently reduce the linear stability analysis to a decoupled set of 4×4 systems of ODEs as in [29]. We give a characterization of the flock solution in the complex plane for (2.20) with $(y_j^0, z_j^0) = (Re^{i\theta_j}, 0)$, where $\theta_j = \frac{2\pi j}{N}$. We consider then the perturbed solution

$$\tilde{y}_j(t) = Re^{i\theta_j}(1 + h_j(t)),$$

with h_j such that $|h_j| \ll 1$ and satisfying

$$\sum_{j=1}^{N} h_j(t) = \sum_{j=1}^{N} h_j'(t) = 0. \tag{2.21}$$

Consider the following relations

$$\tilde{y}_l - \tilde{y}_j = Re^{i\theta_j}\left(e^{\phi_p} h_l - h_j\right),$$
$$|\tilde{y}_l - \tilde{y}_j| \simeq 2R\left|\sin\left(\frac{\phi_p}{2}\right)\right| + \frac{R}{4\left|\sin\left(\frac{\phi_p}{2}\right)\right|}\left[\left(1 - e^{i\phi_p}\right)\left(h_l - \overline{h_j}\right) + \left(1 - e^{-i\phi_p}\right)\left(\overline{h_l} - h_j\right)\right],$$
$$\tilde{z}_l - \tilde{z}_j = Re^{i\theta_j}\left(e^{\phi_p} h_l' - h_j'\right),$$

where $\phi_p = 2\pi(l-j)/N = 2\pi p/N$. We linearize the Cucker-Smale alignment term around the solution up to first order, leading to

$$H(|\tilde{y}_l - \tilde{y}_j|) \simeq H(2R\,|\sin(\phi_p/2)|) + \\ + H'(2R\,|\sin(\phi_p/2)|)\frac{R}{4\left|\sin\left(\frac{\phi_p}{2}\right)\right|}\left[\left(1-e^{i\phi_p}\right)\left(h_l - \overline{h_j}\right) + \left(1-e^{-i\phi_p}\right)\left(\overline{h_l} - h_j\right)\right].$$

Substituting the linearization in (2.20) and neglecting the second order terms, we obtain the following characterization of h''_j

$$\begin{aligned} h''_j =& \frac{1}{N}\sum_{l=1}^{N} H(2R|\sin\phi_p|)\left[e^{i\phi_p}h'_l - h'_j\right] \\ &+ \frac{1}{N}\sum_{\substack{l=1\\ l\neq j}}^{N}\left[G_1(\phi_p/2)(h_j - e^{i\phi_p}h_l) + G_2(\phi_p/2)(\overline{h_l} - e^{i\phi_p}\overline{h_j})\right]. \end{aligned} \tag{2.22}$$

In order to study the behavior of the perturbations h_j, we reduce the complexity of the problem by assuming that h_j satisfies the following relation

$$h_j = \xi_+(t)e^{im\theta_j} + \xi_-(t)e^{-im\theta_j}, \quad h'_j = \xi'_+(t)e^{im\theta_j} + \xi'_-(t)e^{-im\theta_j}, \qquad m \in \mathbb{N}.$$

Therefore, we can express h_l in terms of h_j as

$$h_l = \xi_+(t)e^{im\theta_j}e^{im\phi_p} + \xi_-(t)e^{-im\theta_j}e^{-im\phi_p}, \qquad m \in \mathbb{N}.$$

Inserting the previous expressions in (2.22) and gathering terms in $e^{i\theta_j m}$ and $e^{-i\theta_j m}$, we can characterize ξ_+ and ξ_- as

$$\begin{aligned} \xi''_+ =& \frac{1}{N}\sum_{l=1}^{N} H(2R|\sin\phi_p|)\left[e^{i\phi_p(m+1)} - 1\right]\xi'_+ + I_1(m)\xi_+ + I_2(m)\overline{\xi}_-, \\ \overline{\xi}''_- =& \frac{1}{N}\sum_{l=1}^{N} H(2R|\sin\phi_p|)\left[e^{i\phi_p(m-1)} - 1\right]\overline{\xi}_- + I_2(m)\xi_+ + I_1(-m)\overline{\xi}_-, \end{aligned}$$

where I_1 and I_2 are defined in (2.18) and (2.19). Through a simple manipulation of the sum for the linearized Cucker-Smale term, we obtain that the expression

$$\frac{1}{N}\sum_{l\neq j} H(2R|\sin\phi_p|)\left[e^{i\phi_p(m\pm1)} - 1\right] = \\ \frac{1}{N}\sum_{l\neq j} H(2R|\sin\phi_p|)\left[\cos(\phi_p(m\pm1)) - 1\right] + \frac{i}{N}\sum_{l\neq j} H(2R|\sin\phi_p|)\sin(\phi_p(m\pm1)),$$

is real. Actually, $H(2R|\sin\phi_p|)$ and $\sin(\phi_p(m\pm 1))$ are respectively symmetric and antisymmetric with respect to the values of ϕ_p, so the imaginary part vanishes. Recalling the definition of ϕ_p, we conclude

$$\begin{aligned} J_\pm(m) =& \frac{1}{N}\sum_{p=1}^{N} H\left(2R\left|\sin\left(\frac{2\pi p}{N}\right)\right|\right)\left[\cos\left(\frac{2\pi p}{N}(m\pm 1)\right)-1\right] \\ =& -\frac{4}{N}\sum_{p=1}^{N/2} H\left(2R\left|\sin\left(\frac{2\pi p}{N}\right)\right|\right)\left[\sin^2\left(\frac{\pi p}{N}(m\pm 1)\right)\right]. \end{aligned}$$

Therefore, we reduce the stability analysis to the following system

$$\begin{pmatrix}\xi_+'' \\ \\ \overline{\xi}_-''\end{pmatrix} = \underbrace{\begin{pmatrix} I_1(m) & I_2(m) \\ \\ I_2(m) & I_1(-m)\end{pmatrix}}_{M}\begin{pmatrix}\xi_+ \\ \\ \overline{\xi}_-\end{pmatrix} + \underbrace{\begin{pmatrix} J_+(m) & 0 \\ \\ 0 & J_-(m)\end{pmatrix}}_{J}\begin{pmatrix}\xi_+' \\ \\ \overline{\xi}_-'\end{pmatrix}.$$

Taking the conjugate in the second equation and relabeling $\overline{\xi}_-$ with ξ_- as in [29], the previous system with $\eta_\pm = \xi_\pm'$ is equivalent to

$$\frac{d}{dt}\begin{pmatrix}\xi_+\\ \xi_-\\ \eta_+\\ \eta_-\end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ I_1(m) & I_2(m) & J_+(m) & 0 \\ I_2(m) & I_1(-m) & 0 & J_-(m)\end{pmatrix}\begin{pmatrix}\xi_+\\ \xi_-\\ \eta_+\\ \eta_-\end{pmatrix} = \begin{pmatrix} 0 & \mathrm{Id} \\ M & J\end{pmatrix}\begin{pmatrix}\xi_+\\ \xi_-\\ \eta_+\\ \eta_-\end{pmatrix}. \tag{2.23}$$

At this point, we will do a stability analysis based on the eigenvalues of the matrix of the previous system in a similar way as in Section 2.3.1. Since this analysis can be done in full generality, we consider the linearization, now in vector notation, of the Cucker-Smale alignment term

$$-\sum_{l=1}^{N} g\left(|x_j - x_l|\right)(v_j - v_l),$$

where $g(r)$ denotes any strictly positive function. The corresponding stability matrix L_{CS} for the flock reads

$$L_{\mathrm{CS}} = \begin{pmatrix} 0 & \mathrm{Id} \\ \mathbf{M} & -G\end{pmatrix}.$$

As before, $\mathbf{M}$ denotes stability matrix of the first order model. As the alignment term is linear in the velocity, the matrix G acts on $\mathbf{v} = (v_1,\dots,v_N)^T$, $v_j\in\mathbb{R}^2$, according to the relation

$$(G\mathbf{v})_j = \sum_{l=1}^{N} g\left(|x_j - x_l|\right)(v_j - v_l).$$

In particular, if we denote $||v_j - v_l||_2^2 := (v_j - v_l)^*(v_j - v_l)$ then this relation implies that

$$\mathbf{v}^* G \mathbf{v} = \frac{1}{2} \sum_{j,l=1}^{N} g\left(|x_j - x_l|\right) ||v_j - v_l||_2^2.$$

Consequently, G is positive semi-definite and $G\mathbf{v} = \mathbf{0}$ if and only if $\mathbf{v}$ is "constant" in the sense that $v_j \equiv w$ for some fixed $w \in \mathbb{R}^2$. In other words, $\ker(G) = \mathrm{span}\,\{\mathbf{e}_1, \mathbf{e}_2\}$. By translation invariance of the first order model, both $\mathbf{e}_1 \in \ker(\mathbf{M})$ and $\mathbf{e}_2 \in \ker(\mathbf{M})$ as well.

Note that the eigenvalue problem for L_{CS} is again equivalent to the following quadratic eigenvalue problem for $\mathbf{x} \in \mathbb{C}^{2N}$: $\lambda^2 \mathbf{x} + \lambda G \mathbf{x} - \mathbf{M}\mathbf{x} = \mathbf{0}$. Assuming the normalization $\mathbf{x}^*\mathbf{x} = 1$, the quadratic formula then implies that

$$\lambda = \frac{-\mathbf{x}^* G \mathbf{x} \pm \sqrt{(\mathbf{x}^* G \mathbf{x})^2 + 4\mathbf{x}^* \mathbf{M} \mathbf{x}}}{2}.$$

From this relation and the fact that $\ker(G) \subset \ker(\mathbf{M})$ we conclude that $\ker(L_{\mathrm{CS}}) = \ker(\mathbf{M})$, and moreover that $\Re(\lambda) = 0 \quad \Leftrightarrow \quad \lambda = 0$. Furthermore, $\mathbf{e}_1$ and $\mathbf{e}_2$ generate a single generalized eigenvector whereas each remaining $\mathbf{x} \in \ker(\mathbf{M})$ generates no generalized eigenvectors. Indeed, corresponding to each $\mathbf{x} \in \ker(\mathbf{M})$ the system of equations

$$\begin{pmatrix} 0 & \mathrm{Id} \\ \mathbf{M} & -G \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix}.$$

has a solution if and only if $G\mathbf{x} = \mathbf{0}$ and $\mathbf{u} \in \ker(\mathbf{M})$. Additionally, if $\mathbf{x} = \mathbf{e}_i$ then for any $\mathbf{u} \in \ker(\mathbf{M})$ the system of equations

$$\begin{pmatrix} 0 & \mathrm{Id} \\ \mathbf{M} & -G \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{u}} \\ \tilde{\mathbf{w}} \end{pmatrix} = \begin{pmatrix} \mathbf{u} \\ \mathbf{x} \end{pmatrix}$$

has no solutions. This follows by multiplying the second equation by $\mathbf{x}^*$, then using the facts that $\mathbf{e}_i \in \ker(G) \subset \ker(\mathbf{M})$ and the facts that G and $\mathbf{M}$ are symmetric. In other words, if $g(r)$ is any strictly positive function then

$$\det(L_{\mathrm{CS}} - \lambda \mathrm{Id}) = \lambda^{2+\dim(\ker(\mathbf{M}))} p_g(\lambda),$$

for some polynomial $p_g(\lambda)$ that has non-zero roots. Since this equation holds for any strictly positive function $g(r)$, we may follow the proof of Theorem 2.3.1 to conclude that the second order model has an eigenvalue with positive real part if and only if the first order system has a positive eigenvalue. Moreover, the vectors $(\mathbf{e}_i, \mathbf{e}_i)$ for each $i = 1, 2$ furnish generalized eigenvectors with eigenvalue zero, so remark 2.3.1 applies for this model as well. As a summary, we have shown:

Theorem 2.3.2 *(Spectral Equivalence) The linearized second order system* (2.3) *around the flock ring solution has an eigenvalue with positive real part if and only if the linearized first order system around the ring solution has a positive eigenvalue. Moreover, the flock ring solution is unstable for m-mode perturbations for the second order model* (2.3) *if and only if the ring solution is unstable for m-mode perturbations for the first order model* (2.2).

The linear stability analysis of the previous system leads to the characterization of the same stability areas represented in Figure 2.2.

We numerically investigate the behavior of the eigenvalue with the largest real part, $\Re(\lambda_1)$, of the linearized system (2.23) against the increasing value of communication strength γ for $g(r)$ of the form

$$g(r) := \frac{1}{(1+r^2)^\gamma}.$$

In Figure 2.5, as the potential gets more repulsive at the origin, we see the change from stability to instability, and the rate of convergence to equilibrium depending on γ.

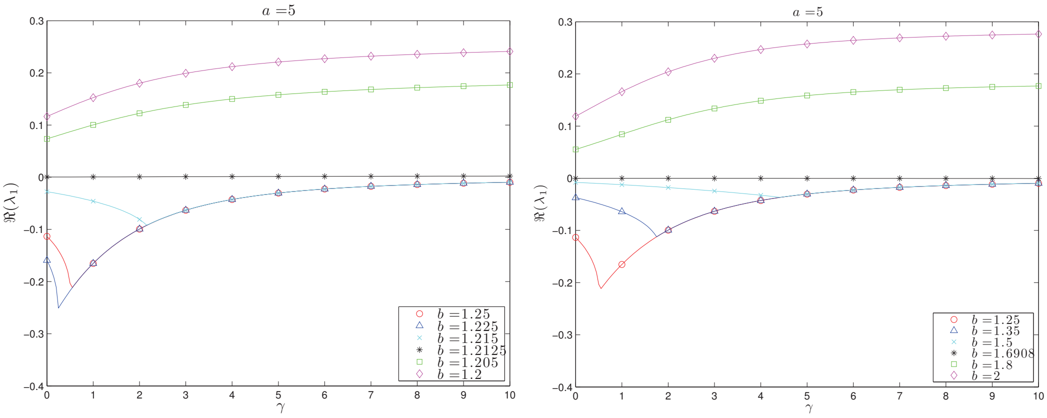


Figure 2.5: The magnitude of $\Re(\lambda_1)$ is influenced by γ, for different values of b and fixed $a = 5$, $N = 10000$.

2.4 Stability for Mill Solutions

This section is meant to complement the results in [29] by analyzing the stability of mill ring solutions with repulsion and attraction. The authors in [29] performed a linear stability analysis on second order models for mill ring solutions. However, they investigated different behaviors of this class of solutions only for attractive potentials. The goal of this section is to explore new behaviors that arise when using repulsive-attractive power law potentials.

2.4.1 Linear Stability Analysis

Let us consider the transformation

$$\begin{cases} y_j(t) = O(t)x_j(t) \\ z_j(t) = O(t)v_j(t) \end{cases}, \quad j = 1, \ldots, N$$

where $O(t)$ is the rotation matrix defined as

$$O(t) = e^{St}, \quad S = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}, \quad \text{and} \quad \dot{O}(t) = Se^{St}.$$

By evaluating $\dot{y}_j(t)$ and $\dot{z}_j(t)$ explicitly, we get after some straightforward computations that

$$\begin{cases} \dot{y}_j(t) = Sy_j(t) + z_j(t) \\ \dot{z}_j(t) = Sz_j(t) + (\alpha - \beta|z_j|^2)z_j(t) + \dfrac{1}{N}\displaystyle\sum_{\substack{l=1 \\ l\neq j}}^{N} \nabla W(y_l - y_j) \end{cases}, \quad j = 1, \ldots, N.$$

A linear stability analysis for mill rings was performed in [29]. Actually, for a fixed number of particles, we have a mill ring solution given by $(y_j^0, z_j^0) = (Re^{i\theta_j}, 0)$, where $\theta_j = \frac{2\pi j}{N}$, and R determined by equation (2.8). With the same notation as in Section 2.3, the analysis in [29] leads to the linear system

$$\begin{pmatrix} \xi'_+ \\ \xi'_- \\ \eta'_+ \\ \eta'_- \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\omega i\alpha + \omega^2 + I_1(m) & -\omega i\alpha + I_2(m) & -\alpha - 2\omega i & \alpha \\ \omega i\alpha + I_2(m) & \omega i\alpha + \omega^2 + I_1(-m) & \alpha & -\alpha + 2\omega i \end{pmatrix} \begin{pmatrix} \xi_+ \\ \xi_- \\ \eta_+ \\ \eta_- \end{pmatrix}. \tag{2.24}$$

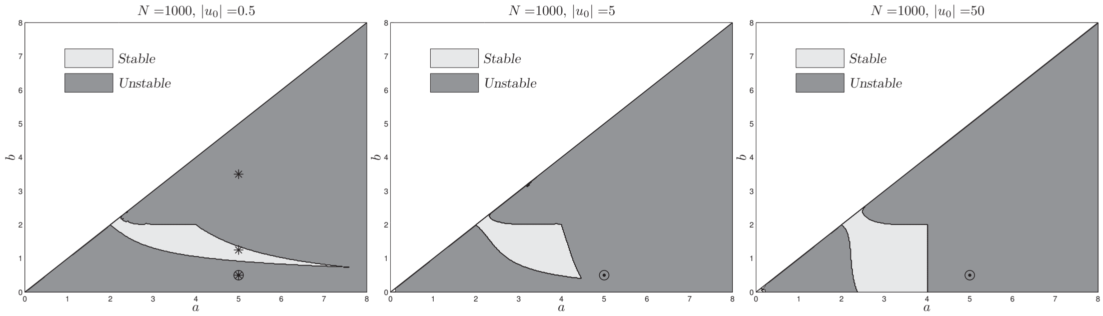
Let us remind that the perturbations are of the form $\tilde{y}_j(t) = Re^{i\theta_j}(1 + h_j(t))$, with $h_j = \xi_+(t)e^{im\theta_j} + \xi_-(t)e^{-im\theta_j}$, $m = 2, 3, \ldots$, such that $|h_j| \ll 1$ and satisfying (2.21), with $(\eta_+, \eta_-) = (\xi'_+, \xi'_-)$. We will make use of (2.24) to study the stability of mill rings with repulsion.

2.4.2 Numerical Tests

Unlike the case of flock solutions where the asymptotic speed does not play role in the linear stability, we will show that the asymptotic speed $|u_0|$ can be used as a bifurcation parameter for mills.

In Table 2.2 we numerically investigate the behavior of the stability region for a fixed number of particles, $N = 1000$, and for increasing values of the asymptotic speed $|u_0|$. We observe that the stability region shrinks with respect to a and gets larger with respect to b. Each stability region in Table 2.2 is computed out of the intersection of the stable areas for the system (2.24) for each perturbation mode $m \geqslant 2$. Note that for $|u_0| = 0$ the stability region coincides with the one for the the first order model (2.2) and for the flock ring solution presented in Figure 2.2.

Table 2.2: Stability region for $N = 1000$ and different values of the asymptotic speed $|u_0|$. Markers ($\odot$) and ($\circledast$) correspond to the explored parameters in Table 2.3 and Table 2.4.

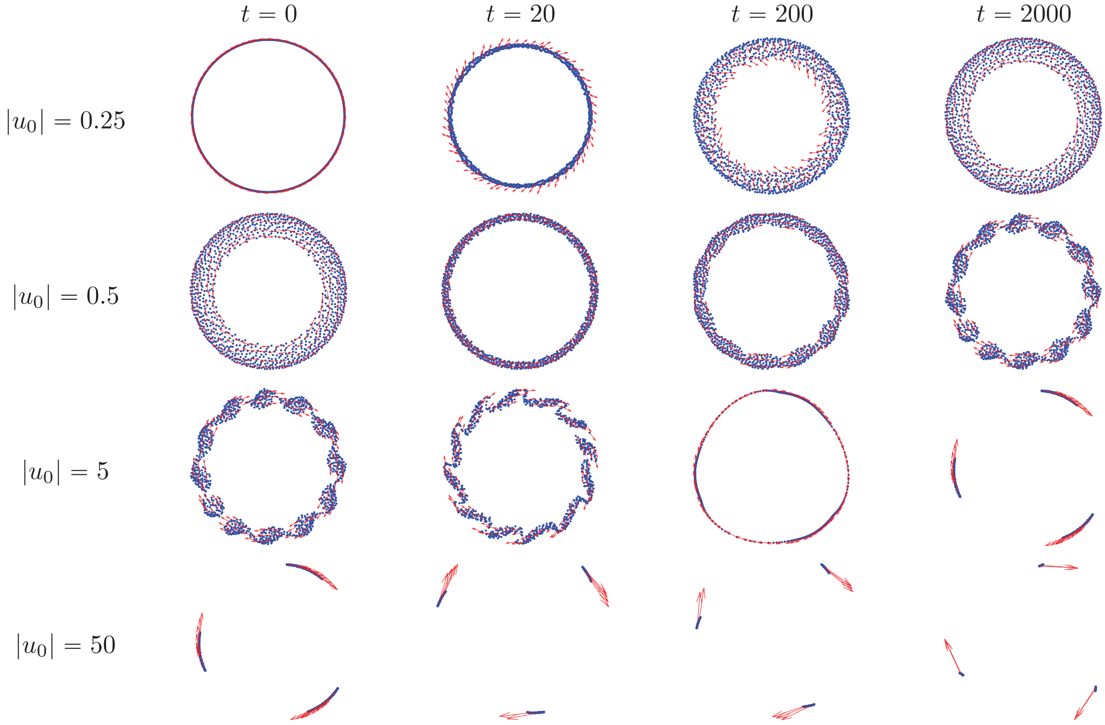


A similar analysis, as done in Subsection 2.3.2, can be performed to study the formation of fat mills and clustered mill solutions. We show how both the fattening and the clustering instability are triggered by tuning the asymptotic speed for a choice of the interaction potential (a and b).

In the case of flock ring solutions we observe cluster solutions or annulus solutions when parameters a and b are chosen respectively “below" or “above" the stability region. In the case of mill solutions, a similar behavior is observed, but this will depend also on the chosen value of $|u_0|$. As an example, we fix $(a, b) = (5, 0.5)$, marked as ($\odot$) in Table 2.2, and we observe the behavioral change of the system for increasing values of the asymptotic speed.

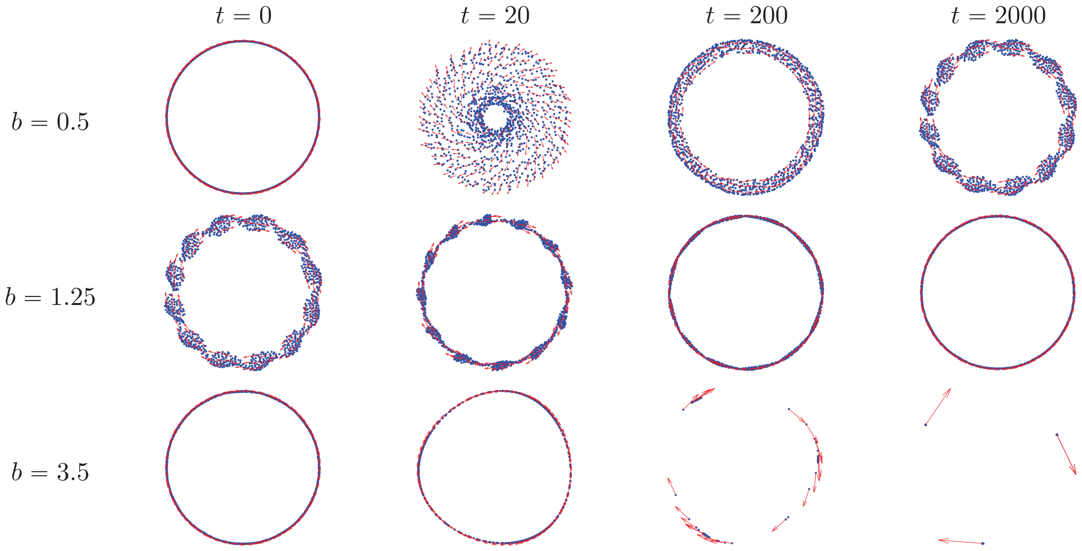
Table 2.3 exhibits this switching behavior from a fat mill to a cluster pattern along with the increment of the asymptotic speed. We observe that for small values of the asymptotic speed fat mill solutions are stable patterns, but when increasing the value of $|u_0|$ the stable solutions form a clustered mill. The first row shows the evolution of the system until stabilization with asymptotic speed $|u_0| = 0.25$ towards an annulus mill. In each of the following rows, we initially start with the previous stable solution (last picture of the preceeding row) and we increase the value of the asymptotic speed. In the second row, we increase its value to $|u_0| = 0.5$ to observe that the stable solution is formed by clusters. The speed in the third row is switched to $|u_0| = 5$ and clusters on lines emerge as a stable configuration. Increasing further the speed to $|u_0| = 50$ in the fourth row, we can observe that clusters on "points" are the stable configuration.

Table 2.3: $N = 1000$ particles, $a = 5, b = 0.5$. The table shows the evolution of a mill ring for increasing values of the speed $|u_0|$. Each row depicts the behavior of the system for a fixed speed, until a stable state is reached. The evolution of the second, third and fourth row is computed starting from the stable pattern of the previous line.



For the sake of completeness, we enrich the analysis by fixing $|u_0| = 0.5$ and considering different values of b, in order to cross the stability region. Therefore in Table 2.4 we show the evolution of a mill ring solution with b taken equal to $0.5, 1.25$ and 3.5, respectively. The parameter choices are marked as $(\circledast)$ in Table 2.2 . The first line of Table 2.4 shows the convergence to the same stable state as the one in second line of Table 2.3, but since the system starts to evolve directly from a ring mill solution the transient behavior is different. Parameters in second line belong to the stability region, see Table 2.2. Therefore, the stable state becomes a mill ring solution. Finally, in the third line we increase b and a three point cluster solution is observed as stable pattern.

Table 2.4: $N = 1000$ particles, $a = 5, |u_0| = 0.5$. The table shows the evolution of a mill ring for increasing values of b, i.e. decreasing repulsion. The evolution of the second and the third row is computed starting from the stable pattern of the previous line.



Mill to Flock and Flock to Mill behavior

We numerically investigate the stability of mill and flock ring solutions for small values of the asymptotic speed, $|u_0|$, and the parameter b, which corresponds to a strong repulsion condition. We perform two representative simulations showing that for a particular choice of the parameters, mill ring solutions can switch to fat flock solutions and conversely flock mill solutions switch to fat mill patterns.

In Table 2.5 we take $N = 100$ particles and we fix $a = 4, b = 0.0005$ and $|u_0| = 0.01$. The frames in the first row show the instability of mill ring solutions for this choice of parameters. The system initially evolves to an almost chaotic state, then particles start to organize and rotate around the center of mass. This rotation actually causes the alignment of the agents and the final fat flock configuration described in the second row.

In Table 2.6 we consider as initial state a flock ring solution. The parameters of the model are $N = 100$, $a = 4$, $b = 0.001$ and $|u_0| = 0.1$. The first row of the table illustrates that the initial configuration is not a stable solution. Therefore, the symmetry of the flock ring is broken and the system exhibits a chaotic behavior. In the second row a rotating dynamic emerges out of the disordered state and finally the system stabilizes to a fat mill solution.

These numerical tests show surprisingly that it is possible to obtain mill con-

Table 2.5: System with $N = 100$ agents, parameters are fixed $a = 4$ and $b = 0.0005$ and $|u_0| = 0.01$. The first row shows that the initial mill ring configuration is unstable. The second row outlines the self organization of the system in a fat flock configuration.

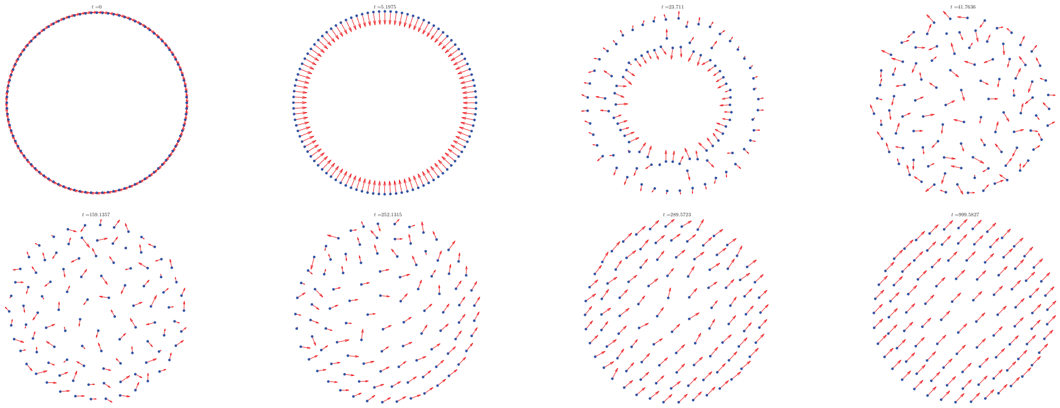
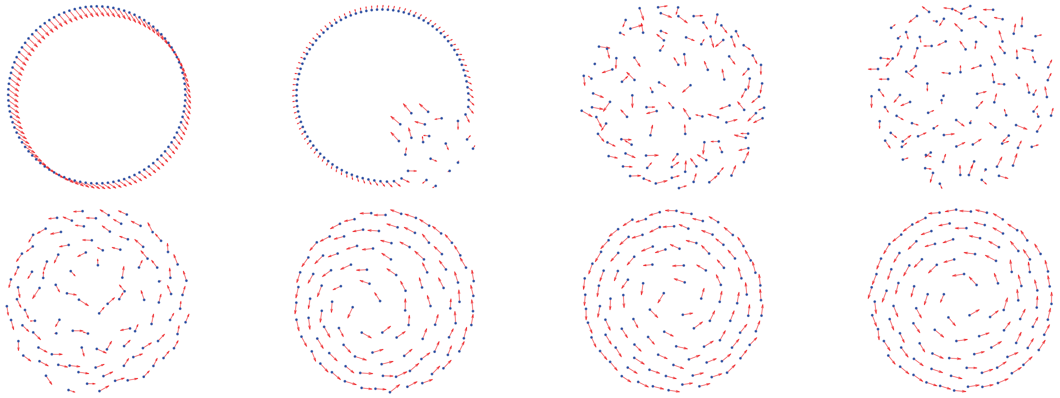


Table 2.6: System with $N = 100$ agents, parameters $a = 4$ and $b = 0.001$ and $|u_0| = 0.1$. The first row shows the instability of the flock ring solution while the second exhibits the convergence to a fat mill type solution.



figurations out of perturbations of initial flock solutions and flock solutions out of perturbations of mill ring solutions. These heteroclinic-kind solutions have not been previously reported. We also remark that the parameter choice is connected to the number of agents we are considering; changing N means finding another set of parameters for which the same switching behavior occurs.

2.5 Conclusions

Numerical simulations of second order models in swarming lead to very rich patterns in a robust way, which indicates their stability. The simplest patterns are flock solutions and mill solutions. We have shown the surprising fact that the spectral stability of flocks for the 2nd order model, either with fixed asymptotic speed (as in (2.1)) or Cucker-Smale alignment (as in (2.3)) terms, is equivalent to the spectral stability for steady states of the first order model. Moreover, particular Fourier mode perturbations allow us to predict the typical instabilities undergone by flock rings. These instabilities, clustering and fattening, are numerically demonstrated. Unlike the case of flock solutions, the stability for mill rings, based on our numerical simulations, cannot be directly related to first order models. Finally, we have shown the numerical instability of mill rings with repulsion in terms of the asymptotic speed and how these instabilities are explained in terms of the linearized analysis of reduced 4 × 4 ODE systems. Some movies illustrating the results in this work can be browsed in giacomoalbi.com/research/simulations/.

CHAPTER 3

Modeling self-organized systems interacting with few individuals

3.1 Introduction

The aim of this paper is to present different levels of description for the dynamic of a large group of agents influenced by a small number of external agent. In a biological context, this corresponds to the behavior of a flock or a school of fishes attacked by one or more predators, or the movement of a heard of sheep guided by a sheepdog. Recently such dynamics have been studied also in robotic research, where scientists tried to control the action of a school of fishes introducing a *fishbot* recognized as leader, see [15].

From the modeling viewpoint this involves considering a microscopic dynamic described by classical flocking models interacting with asset of few individuals characterized in way similar to what was done in [24, 59]. Moreover, motivated by the analysis in [19], we endow the classical dynamic of interaction both with a *metric* as well as a *topological interaction rule*.

Classical flocking models with topological interaction gives more freedom to the model and general results are no more valid, for example in the case of Cucker-Smale model a typical question is under which conditions velocity alignment occurs, in the metric case, when the agents interact all each other. The problem has been solved at the microscopic and kinetic level in [67, 53].

Following the approach in [54] we start from the microscopic dynamic, given by a ODEs system, and we derive two other different levels of description: the mesoscopic (or kinetic) level through a *mean-field limit* and the macroscopic level through a suitable *hydrodynamic approximation*. Here, differently from the first-order macroscopic models proposed in [59], we obtain second-order models for the corresponding continuum dynamic.

Finally we report some numerical examples for the solution of the mean field model in a series of test cases. The simulations have been performed using the fast algorithm recently presented in [6].

3.2 Microscopic model

We are interested in the study of a dynamical system composed of N individuals and N^p external agents with the following general structure

$$\begin{cases} \dfrac{dx_i}{dt} = v_i & \\ & i = 1, \dots, N \\ \dfrac{dv_i}{dt} = \dfrac{1}{N^*} \displaystyle\sum_{j \in \Theta_N^*(x_i)} F(x_i, v_i, x_j, v_j) + \dfrac{1}{N_p} \sum_{k=1}^{N_p} F^p(x_i, v_i, p_k) & \\ \dfrac{dp_h}{dt} = \varphi_h(t, \mathbf{p}, \mathcal{A}\rho^N(p_h, t)) \qquad\qquad h = 1, \dots, N_p & \end{cases} \tag{3.1}$$

where $(\mathbf{x}, \mathbf{v})_i = (x_i, v_i)$ lives in $\mathbb{R}^{2d}$, $d \geqslant 1$, $i = 1, \dots, N$ and $(\mathbf{p})_h = p_h \in \mathbb{R}^{nd}$, with $n = 1, 2$, $h = 1, \dots, N_p$, and $N_p \ll N$.

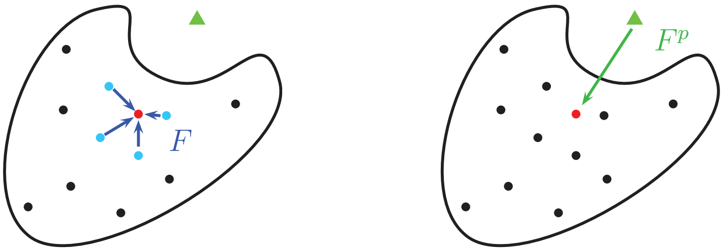


Figure 3.1: Sketch for F and F^p forces acting on each agent of the swarm.

The function F describes the interactions inside the swarm, and F^p depicts the interaction with each external agent p_h. According to the *three zone model* [?], F can be decomposed in

$$F(x_i, v_i, x_j, v_j) = H(x_i, x_j)(v_j - v_i) + A(x_i, x_j) + R(x_i, x_j) + S(v_i)v_i. \tag{3.2}$$

In the above expression H characterizes the *alignment* term, A the *attraction*, R the *repulsion* and S represents a *self propulsion-friction* term. The same decomposition holds for F^p if $p_h = (x_h^p, v_h^p)$, i.e. $n = 2$. In a first order model, $n = 1$, similar interactions can be considered.

Moreover we endow the model with a *topological* rule of interaction. Each agent will interact only with a fix number of agents of their species

$$\Theta_N^*(x_i) = \{\text{the } N^* \text{ closest neighbors respect to } i\}\,. \tag{3.3}$$

A topological interaction is motivated by some recent studies, see [19, 66]. If $N^* = N$ each agent interacts with all the others and the *topological interaction* coincides with the global *metric interaction* [6, 53]. Functions $\varphi_h : [0, +\infty) \times (\mathbb{R}^{nd})^{N_p} \times \mathbb{R} \longrightarrow \mathbb{R}$, describe the evolution in time of each external individual and they depend on the discrete density ρ^N defined as the empirical measure

$$\rho^N(x,t) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i(t)).$$

According to [59] we define $\mathcal{A}$ as a convolution operator $\mathcal{A}\rho^N(x,t) = (\rho^N * \eta)(x,t)$, where η is a smooth kernel with compact support.

Remark 3.2.1 *The* self propelling-friction *term S which appears in (3.2), has the aim to give a desired velocity to the swarm, such velocity is the solution of $S(v) = 0$. In our case, this term is not particularly relevant since each agent will change his velocity mainly according to the action of the leaders/predators rather then trying to reach a desired velocity.*

3.2.1 Classical swarming models

The classical swarming models take into account a global interaction between the agents, that corresponds in our case to choose $N^* = N$ neighbors.

Cucker-Smale model It describes an *alignment dynamic* with

$$H(x_i, x_j)(v_j - v_i) = H(|x_i - x_j|)(v_j - v_i), \tag{3.4}$$

where $H(|x_i - x_j|)$ is a function that measures the strength of the interaction between individuals i and j , and depends on the mutual distance. Under the assumption that closer individuals have more influence than the far distant ones this function is defined as

$$H(r) = \frac{1}{(1+r^2)^\gamma},$$

where $\gamma \geqslant 0$ discriminates the behavior of the solution. We refer to [53, 54] for further details. The classical *Cucker-Smale* model take in account a global interaction between the agents, that corresponds in our case to the choice of $N^* = N$ neighbors in the *topological* rule, (3.3). In this case the standard Cucker-Smale model prescribes perfectly symmetric interactions, as a result total momentum is preserved by the dynamics.

D'Orsogna, Bertozzi et al. model It considers a *self-propelling, attraction and repulsion dynamic* with

$$A(x_i, x_j) + R(x_i, x_j) + S(v_i)v_i = -\nabla_{x_i} W(|x_j - x_i|) + (\alpha - \beta|v_i|^2)v_i \tag{3.5}$$

where $W : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a given potential modeling the short-range repulsion and long-range attraction and α, β are positive parameters. A possible choice is given by the following power law

$$W(r) = \frac{r^a}{a} - \frac{r^b}{b},$$

where $a > b > 0$ are positive parameters. See [80] for more details.

Both the models take in account symmetric interactions between agents, which correspond to the conservation of momentum. Clearly this assumption sounds not very realistic if we want to model interactions among a group of animals. However, the introduction into the models of other features, like the notion of *perception cone* [6, 53] or the concept of *relative distance* [138] breaks the interaction symmetry and consequently loses the conservation of momentum. Note that also the *topological interaction* breaks the symmetry. Other choices have been taken in account as a class of *Quasi-Morse potentials*, [52], or other *power law potentials*.

3.3 Kinetic model

To obtain a mesoscopic description of the system (3.1) we proceed formally through a *mean-field* limit following [53, 54]. Similarly er can recover the same mean-field kinetic equation by deriving first its corresponding Boltzmann model and then considering the asymptotic quasi-invariant limit. Note that, since the limit is done only for the first set of equations, which describes the evolution of the swarm, as result we obtain an hybrid model composed of one kinetic equation and the system of ODEs governing the external agents.

The basic idea of the *mean-field* limit is to derive, through a weak argument, a single evolutionary equation for f^N, the empirical measures defined as

$$f^N(x, v, t) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i(t))\delta(v - v_i(t)).$$

Next, one computes the limit $f^N \longrightarrow f$ for $N \longrightarrow \infty$ and performs a rigorous derivation of the limiting kinetic equation. A well-posedness theory for this asymptotic derivation has been developed in [47] for a general set of swarming models, we refer to it for technical details.

Moreover we assume that for all N the ratio N^*/N is fixed and equal to some constant λ. This supposition allow us to define the *topological density* as $\rho^* = \lambda\bar{\rho}$, where $\bar{\rho} = \int_{\mathbb{R}^{2d}} f(y,v,t)\ dy\ dv$ is constant in time. We assume that, in the mean-field limit, an analogue of the topological set of interaction $\Theta^*_N(x,t)$ is described by a characteristic function on the ball $\mathcal{B}(x,R^*)$, with center x and radius $R^*(x,t)$ such that we have

$$R^*(x,t) = \min\left\{R\ s.t.\ R > 0,\ \int_{\mathbb{R}^d}\int_{\mathcal{B}(x,R)} f(y,v,t)\,dy\,dv = \rho^*\right\}. \tag{3.6}$$

3.3.1 Formal computations of the mean-field limit

We report here the formal computations for the derivation of the mean-field limit. Let us consider a test function $\phi \in \mathcal{C}^1_0(\mathbb{R}^{2d})$ and we compute

$$\frac{d}{dt}\langle f^N(t),\phi\rangle = \frac{1}{N}\sum_{i=1}^N \frac{d}{dt}\phi(x_i(t),v_i(t)) =$$

$$\underbrace{\frac{1}{N}\sum_{i=1}^N \nabla_x\phi(x_i(t),v_i(t))\cdot v_i(t)}_{I_1} + \underbrace{\frac{1}{NN^*}\sum_{i=1}^N\sum_{j\in\Theta^*_N(x_i)} \nabla_v\phi(x_i(t),v_i(t))F(x_i,x_j,v_i,v_j)}_{I_2}$$

$$+\underbrace{\frac{1}{NN_p}\sum_{i=1}^N\sum_{k=1}^{N_p}\nabla_v\phi(x_i(t),v_i(t))F^p(x_i,x^p_k,v_i,v^p_k)}_{I_3}.$$

We solve term by term the summation

$$I_1 = \frac{1}{N}\sum_{i=1}^N \nabla_x\phi(x_i(t),v_i(t))\cdot v_i(t) = \langle f^N(t),\nabla_x\phi\cdot v\rangle$$

Next we rewrite in I_2 the sum on j as the sum on all the agents, where each component of the force field, $F(x,y,v,w)$, is multiplied by $\chi_*(x)$, defined as the characteristic function of the smallest ball centered in x, which contains the topological set $\Theta^*_N(x)$, namely $\chi_*(x) = \chi_{\mathcal{B}(x,R^*)}(x)$, with R^* defined in (3.6). Moreover we consider F as the sum of the Cucker-Smale and the D'Orsogna Bertozzi et al. models, as follows

$$
\begin{aligned}
I_2 = \frac{N}{N^*}\frac{1}{N^2}\sum_{i,j=1}^{N} \nabla_v\phi(x_i(t),v_i(t))F(x_i,x_j,v_i,v_j)\chi_*(x_i) = \\
\frac{1}{\lambda}\frac{1}{N^2}\sum_{i,j=1}^{N} H(|x_i(t)-x_j(t)|)\chi_*(x_i)\nabla_v\phi(x_i(t),v_i(t))\cdot(v_j(t)-v_i(t))+ \\
-\frac{1}{\lambda}\frac{1}{N^2}\sum_{i,j=1}^{N} \nabla_v\phi(x_i(t),v_i(t))\cdot\left(\nabla_{x_i}W(|x_i(t)-x_j(t)|)\right)\chi_*(x_i) = \\
\langle f^N(t),\frac{1}{\lambda}\frac{1}{N}\sum_{j=1}^{N} H(|x-x_j(t)|)\chi_*(x)\nabla_v\phi\cdot(v_j(t)-v)\rangle+ \\
-\langle f^N(t),\frac{1}{\lambda}\frac{1}{N}\sum_{j=1}^{N}\nabla_v\phi\cdot\left(\nabla_x W(|x-x_j(t)|)\right)\chi_*(x)\rangle.
\end{aligned}
$$

Defining the following quantities

$$
\begin{aligned}
\rho^N(x,t) &:= \int_{\mathbb{R}^d} f^N(x,v,t)dv = \frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i(t)), \\
m^N(x,t) &:= \int_{\mathbb{R}^d} vf^N(x,v,t)dv = \frac{1}{N}\sum_{i=1}^{N}v_i(t)\delta(x-x_i(t)),
\end{aligned}
$$

we can express the previous relation in the following way

$$
\begin{aligned}
I_2 = & \left\langle f^N(t),\nabla_v\phi\cdot(H\chi_*)\star m^N\right\rangle - \left\langle f^N(t),\nabla_v\phi\cdot(H\chi_*)\star\rho^N\right\rangle \\
& - \left\langle f^N(t),\nabla_v\phi\cdot((\nabla_x W)\chi_*)\star\rho^N\right\rangle.
\end{aligned}
$$

The last term I_3 has exactly the same structure of the previous one, and we assume that the interaction is led by the same dynamic but with different function values

$$I_3 = \frac{1}{NN_p}\sum_{i=1}^{N}\sum_{k=1}^{N_p}\nabla_v\phi(x_i(t),v_i(t))F^p(x_i,x_k^p,v_i,v_k^p) =$$
$$\frac{1}{NN_p}\sum_{i=1}^{N}\sum_{k=1}^{N_p}\tilde{H}(|x_i(t)-x_k^p(t)|)\nabla_v\phi(x_i(t),v_i(t))\cdot(v_k(t)-v_i(t))+$$
$$-\frac{1}{NN_p}\sum_{i=1}^{N}\sum_{k=1}^{N_p}\nabla_v\phi(x_i(t),v_i(t))\cdot\nabla_{x_i}\tilde{W}(|x_i(t)-x_k^p(t)|) =$$
$$\langle f^N(t),\frac{1}{N_p}\sum_{k=1}^{N_p}\tilde{H}(|x-x_k^p(t)|)\nabla_v\phi\cdot(v_k^p(t)-v)\rangle+$$
$$-\langle f^N(t),\frac{1}{N_p}\sum_{k=1}^{N_p}\nabla_v\phi\cdot\nabla_x\tilde{W}(|x-x_k^p(t)|)\rangle.$$

Analogously to the previous case, we define the following quantities

$$\rho^p(x,t) := \frac{1}{N_p}\sum_{k=1}^{N_p}\delta(x-x_k^p(t)),$$
$$m^p(x,t) := \frac{1}{N_p}\sum_{k=1}^{N_p}v_i(t)\delta(x-x_k^p(t)),$$

we can express the previous relation in the following way

$$I_3 = \langle f^N(t),\nabla_v\phi\cdot\tilde{H}\star m^p\rangle - \langle f^N(t),\nabla_v\phi\cdot\tilde{H}\star\rho^p\rangle - \langle f^N(t),\nabla_v\phi\cdot\nabla_x\tilde{W}\star\rho^p\rangle.$$

Collecting all the terms and integrating by parts in (x,v) we recover the following weak formulation

$$\frac{d}{dt}\langle f^N,\phi\rangle = -\langle v\nabla_x f^N,\phi\rangle - \langle\nabla_v\xi^*(f^N)f^N,\phi\rangle + \langle\zeta^*(f^N)\nabla_v f^N,\phi\rangle+$$
$$-\langle\nabla_v\cdot\tilde{\xi}(f^p)f^N,\phi\rangle + \langle(\nabla_x\tilde{W}\star\rho^p)\nabla_v f^N,\phi\rangle,$$

where

$$\xi^*(f^N)(x,v,t) = \int_{\mathbb{R}^{2d}} H(|x-y|)(v-w)\chi_*(x)f^N(y,w,t)dydw,$$
$$\zeta^*(f^N)(x,v,t) = \int_{\mathbb{R}^d}\nabla_x W(|x-y|)\chi_*(x)\rho^N(y,t)dy,$$

and

$$\tilde{\xi}(f^p)(x,v,t) = \int_{\mathbb{R}^{2d}} \tilde{H}(|x-y|)(v-w)f^p(y,w,t)dydw = \frac{1}{N^p}\sum_{k=1}^{N_p} H(|x-y_k|)(v-w_k),$$

$$\tilde{\zeta}(f^p)(x,v,t) = \int_{\mathbb{R}^d} \nabla_x W(|x-y|)\rho^p(y,t)dy = \frac{1}{N^p}\sum_{k=1}^{N_p} \nabla_x W(|x-y_k|),$$

with

$$f^p(x,v,t) = \frac{1}{N_p}\sum_{k=1}^{N_p} \delta(x - x_k^p(t))\delta(v - v_k^p(t)).$$

Rewriting the main expression we have

$$\begin{aligned}\langle \frac{\partial}{\partial t} f^N + v\cdot\nabla_x f^N + \nabla_v\cdot\xi^*(f^N)f^N - \nabla_v\cdot\zeta^*(f^N)f^N \\ + \nabla_v\cdot\tilde{\xi}(f^p)f^N - \nabla_v\cdot(\nabla_x\tilde{W}\star\rho^p)f^N, \phi\rangle = 0,\end{aligned}$$

and thus the strong form reads

$$\begin{aligned}\frac{\partial}{\partial t} f^N + v\cdot\nabla_x f^N + \nabla_v\cdot\xi^*(f^N)f^N - \nabla_v\cdot\zeta^*(f^N)f^N \\ + \nabla_v\tilde{\xi}(f^p)f^N - (\nabla_x\tilde{W}\star\rho^p)\nabla_v f^N = 0.\end{aligned}$$

Then the limit for $k\to\infty$ of subsequence $(f^{N_k})_k$ leads formally to

$$\partial_t f + v\cdot\nabla_x f = -\nabla_v\cdot[\xi^*(f) + \tilde{\xi}(f^p)]f + (\zeta^*(f) + \tilde{\zeta}(f^p))\cdot\nabla_v f, \tag{3.7}$$

where now

$$\xi^*(f)(x,v,t) = \frac{1}{\lambda}\int_{\mathbb{R}^d}\int_{\mathcal{B}(x,R^*)} H(|x-y|)(v-w)f(y,w,t)dydw,$$

$$\zeta^*(\rho)(x,t) = -\frac{1}{\lambda}\int_{\mathcal{B}(x,R^*)} \nabla_x W(|y-x|)\rho(y,t)dy.$$

Remark 3.3.1 *Note that with the introduction of the characteristic function $\chi_*(x)$, in order to describe the topological interaction, we can not in general use the* well-posedness *theory developed in [47], since it requires the force field to be continuous and locally Lipschitz. This issue could be avoided, with a small change in the initial model: introducing a smooth function $\psi_*^\varepsilon(x)$, corresponding to a regularization of the characteristic function $\chi_*(x)$, as done in [1] for the case of perception cone. Rigorous derivation of the particular mean-field model is under investigation.*

3.3.2 Mesoscopic description

The general mesoscopic model results as a coupled system of a kinetic equation for the swarm and an ODEs for the external agents, as follows

$$\begin{cases} \partial_t f + v \cdot \nabla_x f = -\nabla_v \cdot (\mathcal{E}^*(x,v)f) - \nabla_v \cdot (\mathcal{E}^p(x,v)f), \\ \\ \dot{p}_h = \varphi_h(t, \mathbf{p}, \mathcal{A}\rho(p_h,t)), \qquad\qquad h = 1, \dots, N_p, \end{cases} \tag{3.8}$$

where $\rho(x,t) = \int_{R^d} f(x,v,t)dv$ and

$$\mathcal{E}^*(x,v) = \frac{1}{\lambda}\int_{\mathbb{R}^d}\int_{\mathcal{B}(x,R^*)} F(x,v,y,w)f(y,w)\,dy\,dw, \quad \mathcal{E}^p(x,v) = \frac{1}{N_p}\sum_{k=1}^{N_p} F^p(x,v,p_k). \tag{3.9}$$

Remark 3.3.2 *If we consider the decompositions (3.2) in the case of* Cucker-Smale *and* D'Orsogna-Bertozzi et al. *models and $p_h = (x_h^p, v_h^p)$, we can compute explicitly $\mathcal{E}^*$ and $\mathcal{E}^p$ as*

$$\begin{aligned} \mathcal{E}^*(x,v) = {} & S(v)v + \\ & + \frac{1}{\lambda}\int_{\mathcal{B}(x,R^*)}\left(\int_{\mathbb{R}^d} H(|x-y|)(w-v)f(y,w)dw - \nabla_x W(|x-y|)\rho(y)\right)dy \end{aligned} \tag{3.10}$$

and

$$\mathcal{E}^p(x,v) = \frac{1}{N_p}\sum_{k=1}^{N_p} H^p(|x-x_k^p|)(v_k^p - v) - \frac{1}{N_p}\sum_{k=1}^{N_p}\nabla_x W^p(|x-x_k^p|) \tag{3.11}$$

3.4 Hydrodynamic approximation

Lastly, we will detail a possible macroscopic description of the system. From a numerical point of view, this corresponds to reduce the dimensionality of the problem in such a way that simulations become affordable. Any macroscopic description of a kinetic equation depends upon the local equilibria. In the kinetic theory of rarefied gases, this is a well studied task, which connects Boltzmann equation with the Euler and Navier-Stokes system of fluid dynamics. In the presented situation the determination of the local equilibrium state of the system is, in general, a very difficult task. One usually resorts to approximate equilibrium states which are physically reasonable and simplify the mathematical computations. Here we follow the approach introduced in [58] and subsequently used also in [54, 100].

First let us define the momentum and the temperature of the system as

$$\rho u(x,t) = \int v f dv, \quad T(x,t) = \int |v-u|^2 f dv.$$

In order to obtain a system of equations which describes the evolution of the mass density ρ and the momentum ρu we integrate the kinetic equation in (3.8) against dv and vdv.

According to [58] we impose the closure the momentum assuming that the fluctuations are negligible, i.e., that the temperature $T(x,t) = 0$, and the velocity distribution is *monokinetic*

$$f(x,v,t) = \rho(x,t)\delta(v-u(x,t)).$$

The previous assumptions lead to the following hydrodynamic system

$$\begin{cases} \partial_t \rho + div_x(\rho u) = 0, \\ \partial_t(\rho u) + \nabla_x \cdot (\rho u \otimes u) = \mathcal{F}^*(x,u)\rho(x,t) + \mathcal{F}^p(x,u)\rho(x,t), \\ \dot{p}_h = \varphi_h(t, \mathbf{p}, \mathcal{A}\rho(p_h,t)) \qquad\qquad h = 1,\dots,N_p \end{cases} \tag{3.12}$$

where in the particular case of Cucker-Smale and D'Orsogna Bertozzi et al. model and $p_h = (x_h^p, v_h^p)$ we have

$$\begin{aligned} \mathcal{F}^*(x,u) = {}& S(u)u \\ & + \frac{1}{\lambda}\int_{\mathcal{B}(x,R*)} \left(H(|x-y|)(u(y,t)-u(x,t)) - \nabla_x W(|x-y|)\right)\rho(y,t)\,dy \end{aligned} \tag{3.13}$$

$$\mathcal{F}^p(x,u) = \frac{1}{N_p}\sum_{k=1}^{N_p} H^p(|x-x_k^p|)(v_k^p - u(x,t)) - \frac{1}{N_p}\sum_{k=1}^{N_p} \nabla_x W^p(|x-x_k^p|). \tag{3.14}$$

3.5 Numerical examples

We show here some qualitative simulations of the kinetic model inspired by the work done in [59]. We solve numerically the kinetic model (3.8), using the techniques introduced in Chapter 1.

3.5.1 Confinement: Shepherd dogs

The dynamics considered for the swarm in both cases are characterized by

$$F(x_i, v_i, x_j, v_j) = H(|x_i - x_j|)(v_j - v_i) + \nabla_{x_i} W(|x_i - x_j|),$$

and only a *repulsion* dynamic with respect to p_h in F^p is considered, given by $W^p(r) = -r^c/c$ with $c > 0$, and where $r = |x_i - p_h|$.

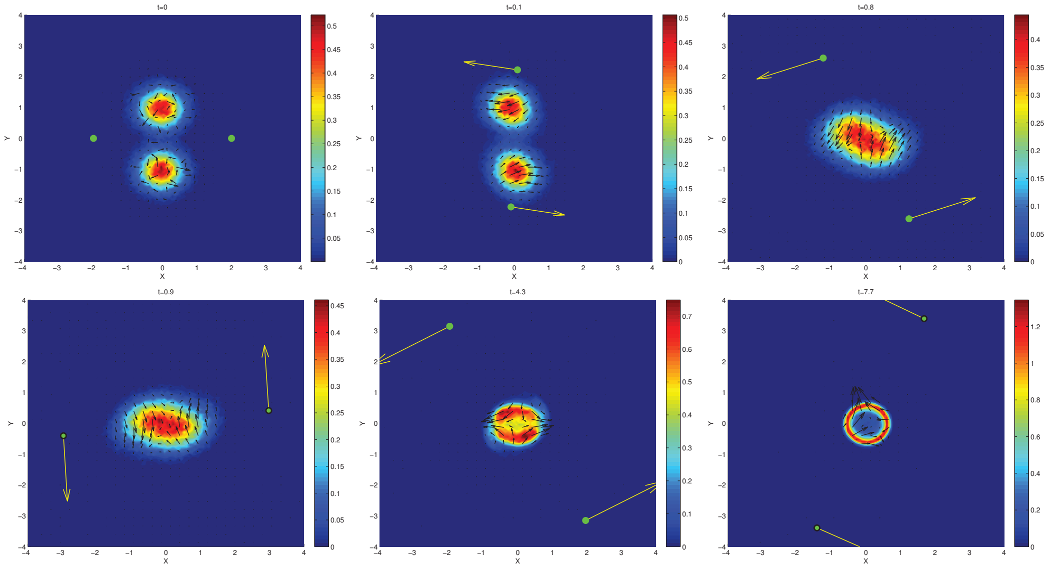


Figure 3.2: Shepherd dogs. With parameters $a = 2.5$, $b = 0.1$, $\gamma = 0.45$ for models (3.4) and (3.5).

The simulation represents the evolution of a swarm controlled by two leaders $p_1, p_2 \in \mathbb{R}^{2n}$, $n = 1$, interacting with the swarm in the following way

$$\varphi_h(t, \mathbf{p}, s_h) = V_p \frac{s_h^\perp}{\sqrt{1 + |s_h|^2}}; \quad V_p = 300, \quad r_p = 5, \quad h = 1, 2.$$

$$\eta(x) = \frac{3}{\pi r_p^6}(\max\{0, r_p^2 - |x|^2\})^2, \quad s_h := \mathcal{A}\rho(x_h^p, t) = (\rho *_x \nabla\eta)(x_h^p, t).$$

As we can see in Figure 3.2 the action of the leaders is able to force the flocking of the particles.

3.5.2 Defense: Swarm attacked by predator

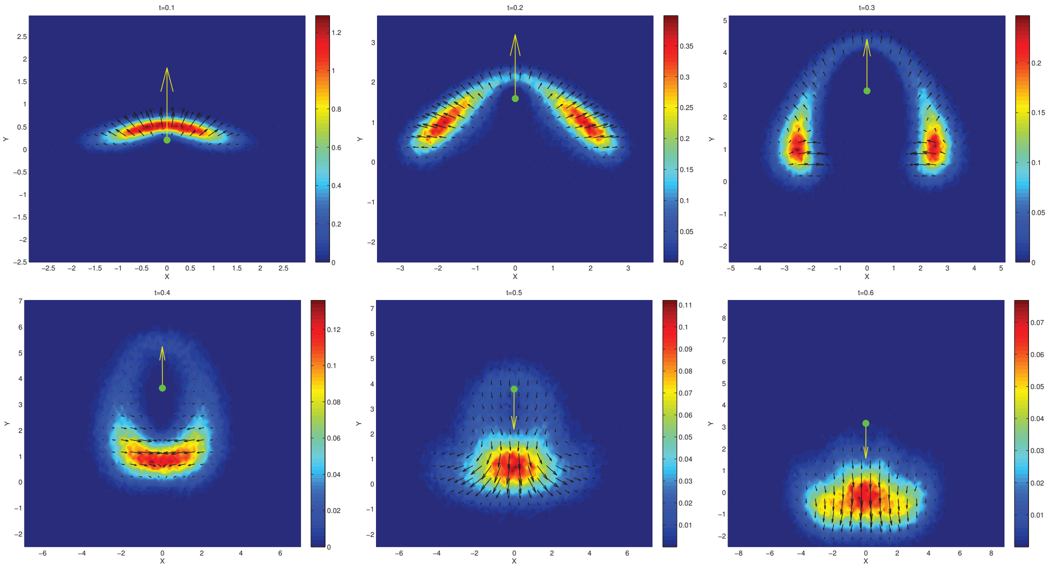


Figure 3.3: Swarm attacked by a predator. With parameters $a = 4$, $b = 2$, $\gamma = 0.45$ for models (3.4) and (3.5).

We consider the evolution of a swarm which undergoes the action of a predator. The predator is modeled by the evolution of $p = (x^p, v^p) \in \mathbb{R}^{2n}$, $n = 2$ and its evolution is lead by the following potential

$$\varphi(t, p, s) = (v^p, V_p s); \quad V_p = 1500, \quad r_p = 5.$$

$$\eta(x) = \frac{3}{\pi r_p^6}(\max\{0, r_p^2 - |x|^2\})^2, \quad s := (\mathcal{A}\rho)(x^p, t) = (\rho * \nabla\eta)(x^p, t).$$

We report the results in Figure 3.3. It is evident how the predator attack splits the flock in two groups which subsequently merge again together.

3.5.3 Followers: Swarm attracted by a leader

We consider the evolution of a swarm which follows the action of a leader. The leader's trajectory is prescribed as a circular trajectory, we report the results in Figure 3.4.

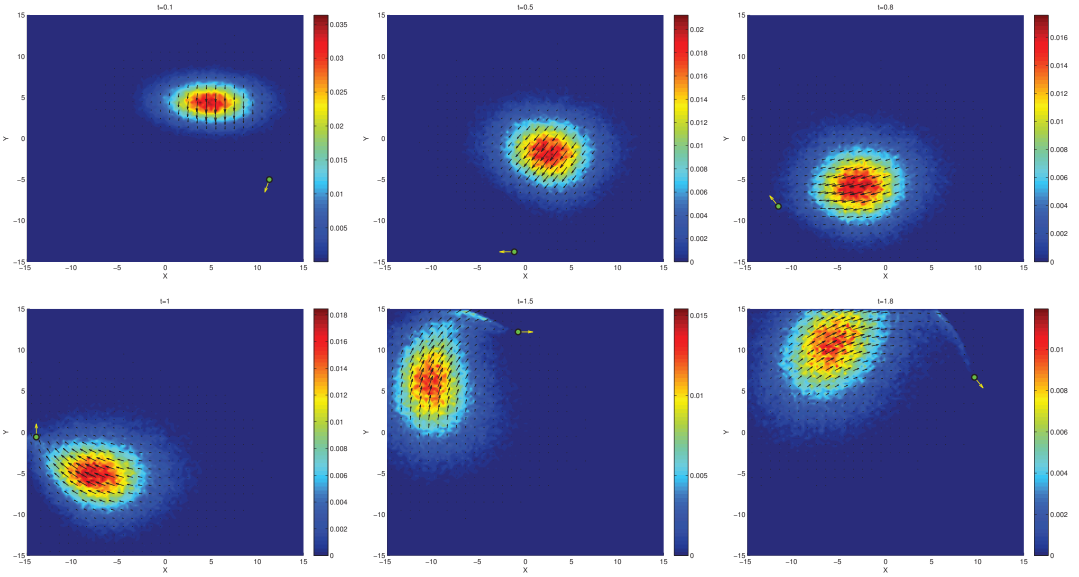


Figure 3.4: Swarm attracted by a leader. With parameters $a = 4$, $b = 2$, $\gamma = 0.45$ for models (3.4) and (3.5).

3.6 Conclusions

In this chapter we focus on the study of self-organized systems interacting with few external individuals, representing a set of leader or predators. We presented different levels of descriptions: (1) *microscopic* as a coupled system of ODEs, describing the evolution of the swarm and its reaction to external individuals. (2) *Mesoscopic*, derived through the mean-field limit, leaving the evolution of the external forces at level of the microscopics, this results as a system of a single kinetic equation of Vlasov-type, coupled with a system of ODEs. (3) Finally we derive the *macroscopic*, assuming the ansatz of a monokinetic distribution, therefore we obtain a system of two PDEs, mass conservation and momentum, coupled with the microscopic evolution of the externals. The whole dynamic is embedded with a *topological interaction*, we show that the equivalent continuous interaction is described by a ball of a radius determined by the local density at point x. Results are validate with several numerical tests. The perspectives of these chapter follow two directions. First at the modeling level, we can see external leaders or predators as a control dynamic on the swarming system, this is further explored in Chapter 4. Second, simulations are performed just for the metric interaction case with the numerical methods developed in Chapter 1, further developments of these Monte Carlo technique are under investigation for the topological interaction.

CHAPTER 4

Kinetic description of optimal control problems and applications to opinion consensus and flocking

4.1 Introduction

The development of mathematical models describing the collective behavior of systems of interacting agents originated a large literature in the recent years with applications to several fields, like biology, engineering, economy and sociology (see [22, 24, 62, 64, 67, 80, 90, 102, 130, 154, 155, 108, 109, 11, 106] and the references therein). Most of these models are at the level of the microscopic dynamic described by a system of ordinary differential equations. Only recently some of these models have been related to partial differential equations through the corresponding kinetic and hydrodynamic description [7, 11, 39, 71, 73, 85, 97, 100, 108, 106, 131, 155]. We refer to the recent surveys in [139, 141, 159] and to the book [146] for an introduction to the subject.

In this paper we consider problems where the collective behavior corresponds to the process of alignment, like in the opinion formation dynamic. Different to the classical approach where individuals are assumed to freely interact with each other, here we are particularly interested in such problems in a constrained setting. We consider feedback type controls for the resulting process and present a kinetic modeling including those controls. This can be used to study the exterior influence of the system dynamics to enforce emergence of non spontaneous desired asymptotic states. Classical examples are given by persuading voters to vote for a specific candidate or by influencing buyers towards a given good or asset [25, 81, 130, 131]. In our model, the external intervention is introduced as an additional control subject to certain bounds, representing the limitations, in terms of economic resources, media availability, etc., of the opinion maker.

Control mechanisms of self-organized systems have been studied for macroscopic models in [60, 61] and for kinetic and hydrodynamic models in [7, 81, 109]. However, in the above references, the control is modeled as a leader dynamics. Therefore, it is given a priori and represented by a supplementary differential model. Also, in [109] the control is modeled a posteriori on the level of the kinetic equation mimicking a classical LQR control approach. Recently, the control of emergent behaviors in multiagent systems has been studied in [49, 88] where the authors develop the idea of sparse optimization (for sparse control it is meant that the policy maker intervenes the minimal amount of times on the minimal amount of individual agents) at the microscopic and kinetic level. We refer also to [27] for results concerning the control of mean-field type systems. Contrary to all those approaches we derive a controller using the model predicitive control framework on the microscopic level and study the related kinetic description for large number of agents. In this way we do not need to prescribe control dynamics a priori or a posteriori but these are obtained automatically based only on the underlying microscopic interactions and a suitable cost functional.

The starting point of our modeling is a general framework which embed several type of collective alignment models. We consider the evolution of N agents where each agent has an opinion $w_i = w_i(t) \in \mathcal{I}$, $\mathcal{I} = [-1, 1]$, $i = 1, \dots, N$ and this opinion can change over time according to

$$\dot{w}_i = \frac{1}{N} \sum_{j=1}^{N} P(w_i, w_j)(w_j - w_i) + u, \qquad\qquad w_i(0) = w_{0i}, \tag{4.1}$$

where the control $u = u(t)$ is given by the minimization of the cost functional over a certain time horizon T

$$u = \text{argmin} \int_0^T \frac{1}{N} \sum_{j=1}^{N} \left(\frac{1}{2}(w_j - w_d)^2 + \frac{\nu}{2} u^2 \right) ds, \qquad u(t) \in [u_L, u_R]. \tag{4.2}$$

In the formulation (4.2) the value w_d is the desired state and $\nu > 0$ is a regularization parameter. We chose a least–square type cost functional for simplicity but other costs can be treated similarly. We additionally prescribe box constraints on the pointwise values of $u(t)$ given by the constants u_L and $u_R > u_L$. The bound constraints on $u(t)$ are required in order to preserve the bounds for w_i. The dynamic in (4.1) describes an average process of alignment between the opinions w_i of the N agents. Typically, the function $P(w, v)$ is such that $0 \leqslant P(w, v) \leqslant 1$ and represents a measure of the inclination of the agents to change their opinion. Usually such function P follows the assumption that extreme opinions are more difficult to be influenced by others [97, 154, 155]. Problem (4.1)-(4.2) may be reformulated as Mayer's problem and solved by Pontryagin's maximum principle [153] or dynamic programming. The main drawback of this approach relies on the fact that the equation for the adjoint variable has to be solved backwards in time over the full time interval $[0, T]$. In particular, for large values of N the computational effort therefore renders the problem unsolvable. Also, an approach $u = \mathcal{P}(x)$ where $\mathcal{P}$ fulfills a Riccati differential equation cannot be pursued here due to the large dimension of $\mathcal{P} \in \mathbb{R}^{N \times N}$ and a possible general nonlinearity in P. This approach is known as LQR controller in the engineering literature [121]. A standard methodology, when dealing with such complex system, is based on model predictive control where instead of solving the above control problem over the whole time horizon, the system is approximated by an iterative solution over a sequence of finite time steps [44, 133, 135].

In order to decrease the complexity of the model when the number of agents is large, a possible approach is to rely on a kinetic description of the process. Along this line of thought, in this work we introduce a Boltzmann model describing the microscopic model in the model predictive control formulation. Moreover, a Fokker-Planck model is derived in the so called quasi-invariant opinion limit. The kinetic models presented in this paper share some common features with the Boltzmann model introduced in [155] in the unconstrained case and with the mean-field constrained models in [49, 88]. Here, however, a remarkable difference with respect to [49, 88] is that, thanks to the receding horizon strategy, the minimization of the cost functional is embedded into the particle interactions. Similarly to [155], this permits to compute explicitly the stationary solutions of the resulting constrained dynamic.

The rest of the manuscript is organized as follows. In the next Section we introduce the model predictive control formulation of system (4.1)-(4.2). In Section 4.3 a binary dynamic corresponding to the constrained system is introduced and a the main properties of the resulting Boltzmann-type kinetic equation are discussed. In particular, estimates for the convergence of the solution towards the desired state are given. Section 4.4 is devoted to the derivation of the Fokker-Planck model and the computation of explicit stationary solutions for the resulting kinetic equation. Some modeling variants are discussed in Section 4.5. In Section 4.6 several numerical results are reported showing the robustness of the present approach. Finally in Section 4.7.2 we extend the results to the case of flocking models, including a mean-field limit for the derivation different of the kinetic approximation of the optimal control problem. We add in the last part some numerical experiments to validate the methodology presented. Some conclusions and future research directions are made at the end of the chapter.

4.2 Model predictive control

In this section we adapt the idea of the moving horizon controller (or instantaneous control) to derive a computable control u at any time t. Compared with the solution to (4.1)-(4.2) this control will in general only be suboptimal. Rigorous results on the properties of u for quadratic cost functional and linear and nonlinear dynamics are available, for example, in [44, 132]. The model predicitive control framework applied here is also called receding horizon strategy or instantaneous control in the engineering literature.

4.2.1 A receding horizon strategy

We consider a receding horizon strategy with horizon of a single time interval. Hence, instead of solving (4.1)-(4.2) on $[0, T]$, we proceeds as follows:

- Split the time interval $[0, T]$ in M time intervals of length Δt and let $t^n = \Delta t\ n$.
- We assume that the control is piecewise constant on time intervals of length $\Delta t > 0$,
$$u(t) = \sum_{n=0}^{M-1} u^n \chi_{[t^n, t^{n+1}]}(t).$$
- Determine the value of the control $u^n \in \mathbb{R}$ by solving for a state $\bar{w}_i$ the (reduced) optimization problem
$$\begin{aligned} &\dot{w}_i = \frac{1}{N}\sum_{j=1}^{N} P(w_i, w_j)(w_j - w_i) + u, && w_i(t^n) = \bar{w}_i, \\ &u^n = \text{argmin}_{u \in \mathbb{R}} \int_{t^n}^{t^{n+1}} \frac{1}{N}\sum_{j=1}^{N} \left(\frac{1}{2}(w_j - w_d)^2 + \frac{\nu}{2}u^2\right) ds, && u \in [u_L, u_R]. \end{aligned} \tag{4.3}$$
- Having the control u^n on the interval $[t^n, t^{n+1}]$, evolve w_i according to the dynamics
$$\dot{w}_i = \frac{1}{N}\sum_{j=1}^{N} P(w_i, w_j)(w_j - w_i) + u^n \tag{4.4}$$
to obtain the new state $\bar{w}_i = w_i(t^{n+1})$.
- We again solve (4.3) to obtain u^{n+1} with the modified initial data.
- Repeat this procedure until we reach $n\Delta t = T$.

The advantage compared with the problem (4.1)-(4.2) is the reduced complexity of (4.3) being an optimization problem in a single real–valued variable u^n. Furthermore, for the quadratic cost and a suitable discretization of (4.4) the solution to (4.3) allows an explicit representation of u^n in terms of $\bar{w}_i$ and $w_i(t^{n+1})$ provided $u_L = -\infty$ and $u_R = \infty$. As shown in section 4.2.2 this allows to reformulate the previous algorithm as a feedback controlled system which in discretized form reads

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{N} \sum_{j=1}^{N} P_{ij}^n (w_j^n - w_i^n) + \Delta t u^n, \qquad w_i^n = \bar{w}_i, \tag{4.5a}$$

$$u^n = -\frac{\Delta t}{\nu N} \sum_{j=1}^{N} (w_j^{n+1} - w_d). \tag{4.5b}$$

Remark 4.2.1 *Later on, bounds on the control u as in (4.3) are required in order to guarantee that opinions $w_i \in \mathfrak{I}$ for all times. Instead of considering the constrained problem (4.3) we will present a condition on ν ensuring this property in the case of a binary interaction model in Proposition 4.3.1 below. This allows to treat (4.3) as an unconstrained problem and does not require to a priori prescribe bounds u_L and u_R. Also note that in general the expression of the control u in terms of w_i^{n+1} and w_d as in equation (4.5b) would be much more involved if the bound constraints u_L, u_R are present.*

4.2.2 Derivation of the feedback controller

We assume for now that $u_L = -\infty$ and $u_R = +\infty$ and assume sufficient regularity conditions such that any minimizer $u \equiv u^n \in \mathbb{R}$ to problem (4.3) fulfills the necessary first order optimality conditions. We further assume that those conditions are also sufficient for optimality and refer to [153] for more details.

The optimality conditions on $[t^n, t^{n+1}]$ and for $\bar{w}_i = w(t^n)_i$ are given by the set of the following equations.

$$\begin{aligned}
\Delta t\ \nu u &= -\frac{1}{N} \sum_{i=1}^{N} \int_{t^n}^{t^{n+1}} \lambda_i dt, \\
\dot{w}_i &= \frac{1}{N} \sum_{j=1}^{N} P(w_i, w_j)(w_j - w_i) + u,\ w_i(t^n) = \bar{w}_i, \\
\dot{\lambda}_i &= -(w_i - w_d) - \frac{1}{N} \sum_{j=1}^{N} R_{ij},\ \lambda_i(t^{n+1}) = 0, \\
R_{ij} &= \lambda_i \partial_{w_i} \{P(w_i, w_j)(w_j - w_i)\} + \lambda_j \partial_{w_j} \{P(w_j, w_i)(w_i - w_j)\}.
\end{aligned}$$

The function $\lambda_i^n = \lambda_i(t)$ is the (Lagrange) multiplier. If we discretize the adjoint equation (backwards in time) by the implicit Euler scheme we obtain due to the boundary conditions

$$\lambda_i^n = -\Delta t\ (w_i^{n+1} - w_d)$$

Further, we may solve for u after discretizing the integral as $\int_{t^n}^{t^{n+1}} f(t)dt = \Delta t\ f^n$ to obtain

$$u = -\frac{\Delta t}{N\nu}\sum_{i=1}^{N}(w_i^{n+1} - w_d)$$

Applying an explicit Euler discretization to the dynamics for w_i on the time interval $[t^n, t^{n+1}]$ and substituting the control we obtained, we observe that the feedback control u is given by (4.5b) and hence the final equation is given by (4.5a).

The previous derivation is obtained by first computing the continuous optimality system and then applying a suitable discretization. However, applying first an explicit Euler discretization and then computing the discrete optimality system leads to the same result. Indeed, consider the discretization of (4.1)–(4.2) in the interval $[t^n, t^{n+1}]$ for constant control u and with $P_{ij}^n = P(w_i^n, w_j^n)$:

$$\begin{aligned} w_i^{n+1} &= w_i^n + \frac{\Delta t}{N}\sum_{j=1}^{N} P_{ij}^n (w_j^n - w_i^n) + \Delta t u, \qquad\qquad w_i^n = \bar{w}_i, \\ u &= \text{argmin}\ \frac{\Delta t}{N}\sum_{j=1}^{N}\left(\frac{1}{2}(w_j^n - w_d)^2 + \frac{\nu}{2}(u^n)^2\right), \end{aligned} \tag{4.6}$$

The discrete Lagrangian is given by

$$\begin{aligned} \mathcal{L}(w,\lambda,u) =& \Delta t\left(\frac{1}{N}\sum_{k=1}^{N}(w_k^{n+1} - w_d)^2 + \frac{\nu}{2}u^2\right) + \frac{1}{N}\sum_{k=1}^{N}\lambda_k^n(w_k^n - \bar{w}_k) \\ &+ \frac{1}{N}\sum_{k=1}^{N}\lambda_k^{n+1}\left(w_k^n - w_k^{n+1} + \frac{\Delta t}{N}\sum_{j=1}^{N} P_{kj}^n (w_j^n - w_k^n) + \Delta t u\right) \end{aligned} \tag{4.7}$$

A minimizer to equation (4.6) fulfills under suitable regularity assumptions the equations (4.6), (4.8) and (4.9).

$$\lambda_i^{n+1} = \lambda_i^n - \Delta t(w_i^n - w_d) - \frac{\Delta t}{N}\sum_{j=1}^{N} R(w_i(t^n), w_j(t^n))\lambda_i^{n+1}, \quad \lambda_i^{n+1} = 0. \tag{4.8}$$

$$0 = \Delta t \nu u^n + \frac{\Delta t}{N}\sum_{j=1}^{N}\lambda_j^{n+1}. \tag{4.9}$$

Upon substituting the terminal condition for λ_j^{n+1} and expressing u in terms of λ_j^{n+1} we obtain the feedback control (4.5b).

Remark 4.2.2 *In order to generalize the idea we may assume that the control acts differently on each agent. For example, one can consider the situation where action of the control u, acting on the single agents, is influenced by the individual opinion. Therefore we replace u in (4.1) by $uQ(w_i)$, where $Q(w)$ is such that $q_m \leqslant Q(w) \leqslant q_M$. Following the previous computation, the action of the control, at discrete time, is driven by*

$$u^n Q_i^n = -\frac{\Delta t}{\nu N}\sum_{j=1}^{N}(w_j^{n+1} - w_d)Q_j^n Q_i^n \tag{4.10}$$

where $Q_i^n = Q(w_i^n)$. Then the control dynamics on the opinion is described by

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{N}\sum_{j=1}^{N} P_{ij}^n(w_j^n - w_i^n) + \Delta t u^n Q_i^n, \qquad w_i^0 \quad = w_{0i}. \tag{4.11}$$

4.3 Boltzmann description of constrained opinion consensus

In this section, we consider a binary Boltzmann dynamic corresponding to the above model predictive control formulation. We emphasize that the assumption that opinions are formed mainly by binary interactions is rather common, see for example [39, 97, 146, 155]. Following [6, 85, 146] the first step is to reduce the dynamic to binary interactions. Let consider the model predictive control system (4.5a)–(4.5b) in the simplified case of only two interacting agents, numbered i and j. Their opinions are modified in the following way

$$\begin{aligned} w_i^{n+1} &= w_i^n + \frac{\Delta t}{2} P_{ij}^n(w_j^n - w_i^n) + \Delta t u^n, \\ w_j^{n+1} &= w_j^n + \frac{\Delta t}{2} P_{ji}^n(w_i^n - w_j^n) + \Delta t u^n, \end{aligned} \tag{4.12}$$

where the control

$$u^n = -\frac{\Delta t}{2\nu}\left((w_j^{n+1} - w_d) + (w_i^{n+1} - w_d)\right), \tag{4.13}$$

is implicitly defined in terms of the opinions pair at the time $n+1$. The above linear system, however, can be easily inverted and its solutions can be written again in the form (4.12) where now the control is expressed explicitly in terms of the opinions pair at time n as

$$u^n = -\frac{1}{2}\frac{\Delta t}{\nu + \Delta t^2}\left((w_j^n - w_d) + (w_i^n - w_d)\right) - \frac{1}{2}\frac{\Delta t^2}{\nu + \Delta t^2}\left(P_{ij} - P_{ji}\right)(w_j^n - w_i^n). \tag{4.14}$$

Note that, as a result of the inversion of the 2×2 matrix characterizing the linear system (4.12)-(4.13), in the explicit formulation the control contains a term of order Δt^2.

4.3.1 Binary interaction models

In order to derive a kinetic equation we introduce a density distribution of particles $f(w,t)$ depending on the opinion variable $w \in \mathcal{I}$ and time $t \geqslant 0$. The precise meaning of the density f is the following. Given the population of agents under study, if the opinions are defined on a subdomain $\Omega \subset \mathcal{I}$, the integral

$$\int_\Omega f(w,t)\, dw$$

represents the number density of individuals with opinion included in Ω at time $t > 0$. It is assumed that the density function is normalized to 1, that is

$$\int_\mathcal{I} f(w,t)\, dw = 1.$$

The kinetic model can be derived by considering the change in time of $f(w,t)$ depending on the interactions with the other individuals. This change depends on the balance between the gain and loss due to the binary interactions.

Accordingly to the explicit binary interaction (4.12), two agents with opinion w and v modify their opinion as

$$\begin{aligned} w^* &= (1-\alpha P(w,v))\, w + \alpha P(w,v) v - \frac{\beta}{2}\left((v-w_d)+(w-w_d)\right) \\ &\quad - \alpha\frac{\beta}{2}((P(w,v)-P(v,w))(w-v)) + \Theta_1 D(w), \\ v^* &= (1-\alpha P(v,w))\, v + \alpha P(v,w) w - \frac{\beta}{2}\left((v-w_d)+(w-w_d)\right) \\ &\quad - \alpha\frac{\beta}{2}((P(v,w)-P(w,v))(v-w)) + \Theta_2 D(w), \end{aligned} \tag{4.15}$$

where we included an additional noise term as in [155], to take into account effects falling outside the description of the model, like changes of opinion due to personal access to information. In (4.15) we defined the following nonnegative quantities

$$\alpha = \frac{\Delta t}{2}, \qquad \beta = \frac{4\alpha^2}{\nu + 4\alpha^2}, \tag{4.16}$$

which represent the strength of the compromise and of the control respectively. The noise term is characterized by the random variables Θ_1 and Θ_2 taking values on a set $\mathcal{B} \subset \mathbb{R}$, with identical distribution of mean zero and variance σ^2 measuring the the degree of spreading of opinion due to diffusion. The function $D(\cdot)$ represents the local relevance of diffusion for a given opinion, and is such that $0 \leqslant D(w) \leqslant 1$.

In the absence of diffusion, from (4.15) it follows that

$$\begin{aligned} w^* + v^* &= (1-\beta)\,(v+w) + 2\beta w_d + \alpha(1-\beta)(P(w,v) - P(v,w))(v-w) && \text{(4.17a)} \\ w^* - v^* &= (w-v)(1-\alpha(P(w,v)+P(v,w))), && \text{(4.17b)} \end{aligned}$$

thus in general the mean opinion is not conserved. Since $0 \leqslant P(w,v) \leqslant 1$, if we assume $0 \leqslant \alpha \leqslant 1/2$ from (4.17b) we have

$$|w^* - v^*| = (1 - \alpha(P(w,v) + P(v,w))\,|w-v| \leqslant (1-2\alpha)|w-v|, \tag{4.18}$$

which tells that the relative distance in opinion between two agents cannot increase after each interaction.

When dealing with a kinetic problem in which the variable belongs to a bounded domain we must deal with additional mathematical difficulties in the definition of agents interactions. In fact, it is essential to consider only interactions that do not produce values outside the finite interval. The following proposition gives a sufficient condition to preserve the bounds.

Proposition 4.3.1 *Let us assume that* $0 < P(w,v) \leqslant 1$ *and*

$$\frac{\beta}{2} \leqslant \alpha p, \qquad |\Theta_i| < d\left(1 - \frac{\beta}{2}\right), \quad i = 1,2 \tag{4.19}$$

where $p = \min_{w,v\in\mathcal{I}}\{P(w,v)\} > 0$ *and* $d = \min_{w\in\mathcal{I}}\{(1-w)/D(w), D(w) \neq 0\} > 0$, *then the binary interaction* (4.15) *preserves the bounds, i.e. the post-interaction opinions* w^*, v^* *are contained in* $\mathcal{I} = [-1,1]$.

Proof 4.3.1 *We will proceed in two subsequent steps, first by considering the case of interactions without noise and second by including the noise action. Let us define the following quantity*

$$\gamma = \alpha\left(1 - \frac{\beta}{2}\right)P(w,v) + \alpha\frac{\beta}{2}P(v,w), \tag{4.20}$$

where $0 \leqslant \beta \leqslant 1/2$ *by definition.*
Thus relation (4.15) *in absence of noise can be rewritten as*

$$w^* = \left(1 - \gamma - \frac{\beta}{2}\right)w + \left(\gamma - \frac{\beta}{2}\right)v + \beta w_d, \tag{4.21}$$

therefore it is sufficient that the following bounds are satisfied

$$\frac{\beta}{2} \leqslant \gamma \leqslant 1 - \frac{\beta}{2} \tag{4.22}$$

to have a convex combination of w, v and w_d. From equation (4.20), by the assumption on $P(w,v)$, we have $\alpha p \leqslant \gamma \leqslant \alpha$. Therefore the left bound requires that $\alpha p \leqslant \beta/2$, which gives the first assumption in (4.19).

If we now consider the presence of noise, we have

$$w^* = \left(1 - \gamma - \frac{\beta}{2}\right) w + \left(\gamma - \frac{\beta}{2}\right) v + \beta w_d + D(w)\Theta_1. \tag{4.23}$$

Equation (4.23) implies the following inequalities

$$\begin{aligned} w^* &\leqslant \left(1 - \gamma - \frac{\beta}{2}\right) w + \left(\gamma - \frac{\beta}{2}\right) + \beta w_d + D(w)\Theta_1 \\ &\leqslant \left(1 - \gamma - \frac{\beta}{2}\right) w + \left(\gamma + \frac{\beta}{2}\right) + D(w)\Theta_1. \end{aligned}$$

Finally, the last relation is bounded by one if

$$\Theta_1 \leqslant \left(1 - \gamma - \frac{\beta}{2}\right) \frac{(1-w)}{D(w)}, \qquad D(w) \neq 0.$$

which yields the second condition in (4.19). The same results are readily obtained for the post interacting opinion v^*.

Remark 4.3.1 *From the above proposition it is clear that agents should have a minimal amount of propensity to change their opinion in order for the control to act without risking to violate the opinion bounds. This reflects the fact that extreme opinions are very difficult to change and cannot be controlled in general without some additional assumption or model modification. In the case of $\Theta_i = 0$, $\alpha \neq 0$ we obtain from (4.19) the condition*

$$\frac{2\alpha}{\nu + 4\alpha^2} \leqslant p.$$

This condition can be satisfied provided either α is sufficiently small or ν sufficiently large.

4.3.2 Main properties of the Boltzmann description

In general we can recover the time evolution of the density $f(w,t)$ through (4.15) considering for a suitable test function $\varphi(w)$ an integro-differential equation of Boltzmann type in weak form [146]

$$\frac{d}{dt}\int_{\mathcal{I}} \varphi(w)f(w,t)dw = (Q(f,f),\varphi), \tag{4.24}$$

where

$$(Q(f,f),\varphi) = \left\langle \int_{\mathcal{I}^2} B_{int}\left(\varphi(w^*) - \varphi(w)\right) f(w,t)f(v,t)\ dw\ dv \right\rangle. \tag{4.25}$$

In (4.25), as usual, $\langle\,\cdot\,\rangle$ denotes the expectation with respect to the random variables Θ_i, $i = 1, 2$ and the nonnegative interaction kernel B_{int} is related to the probability of the microscopic interactions. The simplest choice which assures that the post interacting opinions preserves the bounds is given by

$$B_{int} = B_{int}(w, v, \Theta_1, \Theta_2) = \eta\chi(|w^*| \leqslant 1)\chi(|v^*| \leqslant 1) \tag{4.26}$$

where $\eta > 0$ is a constant rate and $\chi(\,\cdot\,)$ is the indicator function. A main simplification occurs if the bounds of w^*, v^* are preserved by (4.15) itself and the interaction kernel is independent on w, v, this will corresponds the classical Boltzmann equation for Maxwell molecules. In the rest of the paper, thanks to Proposition 4.3.1, we will pursue this direction. Following the derivation in [62, 146] the present results can be extended to kernels in the form (4.26).

Let us assume that $|w^*| \leqslant 1$ and $|v^*| \leqslant 1$, therefore the interaction dynamic of $f(w,t)$ can be described by the following Boltzmann operator

$$(Q(f,f),\varphi) = \eta\left\langle \int_{\mathcal{I}^2} \left(\varphi(w^*) - \varphi(w)\right) f(w,t)f(v,t)\ dw\ dv \right\rangle. \tag{4.27}$$

The above collisional operator guarantees the conservation of the total number of agents, corresponding to $\varphi(w) = 1$, which is the only conserved quantity of the process. Let us remark that, since $f(w,t)$ is compactly supported in $\mathcal{I}$ then by conservation of the moment of order zero all the moments are bounded. By the same arguments in [155] the existence of a uniform bound on moments implies that the class of probability densities $\{f(w,t)\}_{t\geqslant 0}$ is tight, so that any sequence $\{f(w,t_n)\}_{t_n\geqslant 0}$ contains an infinite subsequence which converges weakly as $t \to \infty$ to some probability measure f_∞.

For $\varphi(w) = w$, we obtain the evolution of the average opinion. We have

$$\frac{d}{dt}\int_{\mathcal{I}} wf(w,t)dw = \eta\left\langle \int_{\mathcal{I}^2} (w^* - w)\, f(w,t)f(v,t)\ dw\ dv \right\rangle \tag{4.28}$$

or equivalently

$$\frac{d}{dt}\int_{\mathcal{I}} wf(w,t)dw = \frac{\eta}{2}\left\langle \int_{\mathcal{I}^2} (w^* + v^* - w - v)\, f(w,t)f(v,t)\; dw\; dv \right\rangle. \tag{4.29}$$

Indicating the average opinion as

$$m(t) = \int_{\mathcal{I}} wf(w,t)\, dw, \tag{4.30}$$

from relation (4.29) and (4.17a), since Θ_i, $i = 1, 2$ have zero mean, we obtain

$$\begin{aligned}\frac{d}{dt}m(t) &= \frac{\eta}{2}\beta \int_{\mathcal{I}^2} (2w_d - w - v)\, f(v)f(w)\; dw\, dv + \\ &\quad + \frac{\eta}{2}\alpha(1-\beta)\int_{\mathcal{I}^2} (P(w,v) - P(v,w))\,(v-w)f(v)f(w)\; dw\, dv \\ &= \eta\beta(w_d - m(t)) + \eta\alpha(1-\beta)\int_{\mathcal{I}^2} (P(w,v) - P(v,w))vf(v)f(w)\; dw\, dv.\end{aligned} \tag{4.31}$$

Note that the above equation for a general P is not closed. Since $0 \leqslant P(w,v) \leqslant 1$ we have $|P(w,v) - P(v,w)| \leqslant 1$, then we can bound the derivative

$$\eta\beta w_d - \eta(\beta + \alpha(1-\beta))m(t) \leqslant \; \frac{d}{dt}m(t) \; \leqslant \eta\beta w_d - \eta(\beta - \alpha(1-\beta))m(t)$$

solving on both sides we obtain the following estimate

$$\begin{aligned} m(t) &\geqslant \frac{\beta}{\beta + \alpha(1-\beta)}\left(1 - e^{-\eta(\beta+\alpha(1-\beta))t}\right) w_d + m(0)e^{-\eta(\beta+\alpha(1-\beta))t} \\ m(t) &\leqslant \frac{\beta}{\beta - \alpha(1-\beta)}\left(1 - e^{-\eta(\beta-\alpha(1-\beta))t}\right) w_d + m(0)e^{-\eta(\beta-\alpha(1-\beta))t}. \end{aligned}$$

If we now assume that

$$\nu < 4\alpha, \tag{4.32}$$

then $\beta - \alpha(1-\beta) > 0$ and if the average $m(t) \to m_\infty$ as $t \to \infty$ we have the bounds

$$\frac{4\alpha}{4\alpha + \nu} w_d \leqslant m_\infty \leqslant \frac{4\alpha}{4\alpha - \nu} w_d. \tag{4.33}$$

Therefore small values of ν force the mean opinion towards the desired state. In the symmetric case $P(v,w) = P(w,v)$, equation (4.31) is in closed form and can be solved explicitly

$$m(t) = \left(1 - e^{-\eta\beta t}\right) w_d + m(0)e^{-\eta\beta t} \tag{4.34}$$

which in the limit $t \to \infty$ converges to w_d, for any choice of the control parameters.

Let us now consider the case $\varphi(w) = w^2$ in the simplified situation of $P(w,v) = 1$. We have

$$\frac{d}{dt}\int_{\mathcal{I}} w^2 f(w,t)dw = \frac{\eta}{2}\left\langle \int_{\mathcal{I}^2} \left((w^*)^2 + (v^*)^2 - w^2 - v^2\right) f(w,t)f(v,t)\ dw\ dv \right\rangle. \tag{4.35}$$

Denoting by

$$E(t) = \int_{\mathcal{I}} w^2 f(w,t)dw, \tag{4.36}$$

easy computations show that

$$\begin{aligned}\frac{d}{dt}E(t) =& -\eta\left(2\alpha(1-\alpha) + \beta\left(1 - \frac{\beta}{2}\right)\right)(E(t) - m(t)^2) - 2\eta\beta\left(\beta(m(t)^2 - w_d^2)\right.\\ &\left.+(1-\beta)m(t)(m(t) - w_d)\right) + \eta\sigma^2\int_{\mathcal{I}} D(w)f(w,t)\,dw,\end{aligned} \tag{4.37}$$

where we used the fact that Θ_i, $i = 1, 2$ have zero mean and variance σ^2. In absence of diffusion, since $m(t) \to w_d$ as $t \to \infty$, we obtain that $E(t)$ converges exponentially to w_d^2 for large times. Therefore the quantity

$$\int_{\mathcal{I}} f(w,t)(w - w_d)^2\,dv = E(t)^2 + w_d^2 - 2m(t)w_d, \tag{4.38}$$

goes to zero as $t \to \infty$. This shows that, under the above assumptions, the steady state solution has the form of a Dirac delta $f_\infty(w) = \delta(w - w_d)$ centered in the desired opinion state.

4.4 Fokker-Planck modeling

In the general case, it is quite difficult to obtain analytic results on the large time behavior of the kinetic equation (4.27). As it is usual in kinetic theory, particular asymptotic limit of the Boltzmann model result in simplified models, generally of Fokker-Planck type, for which the study of the theoretical properties is often easier [146].

4.4.1 The quasi-invariant opinion limit

The main idea is to rescale the interaction frequency η, the propensity strength α, the diffusion variance σ^2 and the action of the control ν at the same time, in order to maintain at level of the asymptotic procedure the memory of the microscopic interactions (4.15). This approach is usually referred to as quasi–invariant opinion limit [146, 155] and is closely related to the grazing collision limit of the Boltzmann equation for Coulombian interactions (see [89, 160]).

We make the following scaling assumptions

$$\alpha = \varepsilon, \qquad \eta = \frac{1}{\varepsilon}, \qquad \sigma^2 = \varepsilon\varsigma, \qquad \nu = \varepsilon\kappa, \tag{4.39}$$

where $\varepsilon > 0$ and as a consequence the coefficient β in (4.15) takes the form

$$\beta = \frac{4\varepsilon}{\kappa + 4\varepsilon}.$$

This corresponds to the situation where the interaction operator concentrates on binary interactions which produce a very small change in the opinion of the agents. From a modeling viewpoint, we require that scaling (4.39) in the limit $\varepsilon \to 0$ preserves the main macroscopic properties of the kinetic system. To this aim, let us observe that the evolution of the scaled first two moments for $P(w, v) = 1$ reads

$$\begin{aligned}
\frac{d}{dt}m(t) =& \frac{4}{\kappa + 4\varepsilon}(w_d - m(t)),\\
\frac{d}{dt}E(t) =& -2\left((1-\varepsilon) + \frac{2}{\kappa + 4\varepsilon}\left(1 - \frac{2\varepsilon}{\kappa + 4\varepsilon}\right)\right)(E(t) - m(t)^2)\\
&- \frac{8}{\kappa + 4\varepsilon}\left(\frac{4\varepsilon}{\kappa + 4\varepsilon}(m(t)^2 - w_d^2) + \left(1 - \frac{4\varepsilon}{\kappa + 4\varepsilon}\right)m(t)(m(t) - w_d)\right)\\
&+ \varsigma\int_{\mathcal{I}} D(w)f(w,t)\,dw,
\end{aligned}$$

which in the limit $\varepsilon \to 0$ gives

$$\frac{d}{dt}m(t) = \frac{4}{\kappa}(w_d - m(t)), \tag{4.40}$$

$$\begin{aligned}
\frac{d}{dt}E(t) =& -2\left(1 + \frac{2}{\kappa}\right)(E(t) - m(t)^2)\\
&- \frac{8}{\kappa}m(t)(m(t) - w_d) + \varsigma\int_{\mathcal{I}} D(w)f(w,t)\,dw.
\end{aligned} \tag{4.41}$$

This shows that in order to keep the effects of the control and the diffusion in the limit it is essential that both ν and σ^2 scale as ε.

In the sequel we show how this approach leads to a constrained Fokker–Planck equation for the description of the opinion distribution. Even if our computations are formal, following the same arguments in [146, 155] it is possible to give a rigorous mathematical basis to the derivation. Here we omit the details for brevity.

The scaled equation (4.27) reads

$$\frac{d}{dt}\int_{\mathcal{I}}\varphi(w)f(w,t)dw = \frac{1}{\varepsilon}\left\langle \int_{\mathcal{I}^2} \left(\varphi(w^*)-\varphi(w)\right) f(w,t)f(v,t)\ dw\ dv \right\rangle \tag{4.42}$$

where the scaled binary interaction dynamic (4.15) can be written as

$$w^* - w = \varepsilon P(w,v)(v-w) + \frac{2\varepsilon}{\kappa+4\varepsilon}\left(2w_d - (w+v)\right) + \Theta_1^\varepsilon D(w) + O(\varepsilon^2), \tag{4.43}$$

where Θ_1^ε is a random variable with zero mean and variance $\varepsilon\varsigma$.

In order to recover the limit as $\varepsilon \to 0$ we consider the second-order Taylor expansion of φ around w

$$\varphi(w^*)-\varphi(w) = (w^*-w)\varphi'(w) + \frac{1}{2}(w^*-w)^2\varphi''(\tilde{w}) \tag{4.44}$$

where for some $0 \leqslant \vartheta \leqslant 1$,

$$\tilde{w} = \vartheta w^* + (1-\vartheta)w.$$

Therefore, inserting this expansion in the interaction integral (4.42) we get

$$\frac{1}{\varepsilon}\left\langle \int_{\mathcal{I}^2}\left((w^*-w)\,\varphi'(w) + \frac{1}{2}\,(w^*-w)^2\,\varphi''(w)\right) f(w)f(v)\ dwdv \right\rangle + R(\varepsilon). \tag{4.45}$$

The term $R(\varepsilon)$ denotes the remainder and is given by

$$R(\varepsilon) = \frac{1}{2\varepsilon}\left\langle \int_{\mathcal{I}^2} (w^*-w)^2\left(\varphi''(\tilde{w})-\varphi''(w)\right)f(w)f(v)\ dwdv \right\rangle. \tag{4.46}$$

Using now (4.43) we can write

$$\begin{aligned}\frac{1}{\varepsilon}\int_{\mathcal{I}^2}\Bigg[\Bigg(&P(w,v)(v-w) + \frac{2\varepsilon}{\kappa+4\varepsilon}\left(2w_d-(w+v)\right)\Bigg)\varphi'(w)\\ &+\frac{\varsigma}{2}D(w)^2\varphi''(w)\Bigg] f(w)f(v)\ dwdv + R(\varepsilon) + O(\varepsilon),\end{aligned} \tag{4.47}$$

where we used the fact that Θ_1^ε has zero mean and variance $\varepsilon\varsigma$.

By the same arguments in [155] it is possible to show rigorously that (4.46) converges to zero as soon as $\varepsilon \to 0$. Therefore we have as limiting operator of (4.27) the following

$$\begin{aligned}\frac{d}{dt}\int_{\mathcal{I}}\varphi(w)f(w)dw = \int_{\mathcal{I}^2}&\left(P(w,v)(v-w) + \frac{4}{\kappa}\left(w_d - \frac{w+v}{2}\right)\right)\varphi'(w)f(w)f(v)dwdv\\ &+\frac{\varsigma}{2}\int_{\mathcal{I}} D(w)^2\varphi''(w)f(w)\ dw.\end{aligned}$$

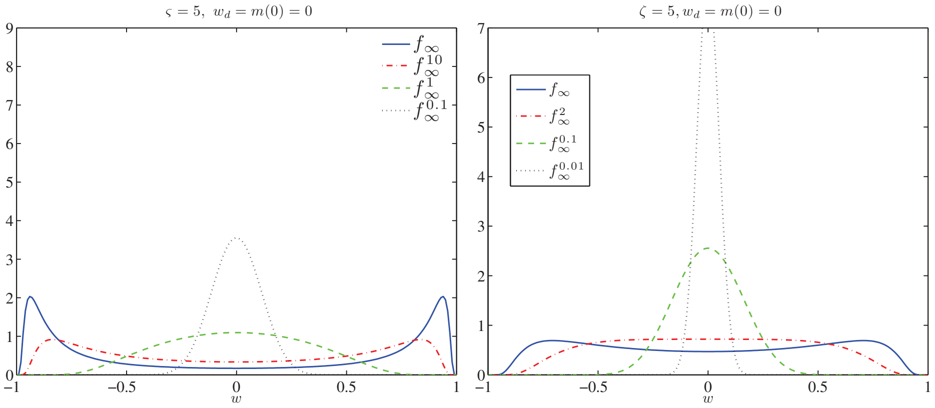


Figure 4.1: Continuous line and dashed lines represent the steady solutions f_∞ and f_∞^κ, respectively. On the left $w_d = m(0) = 0$ with diffusion parameter $\varsigma = 5$, on the right $w_d = m(0) = 0$ with diffusion parameter $\varsigma = 2$. In both cases the steady solution changes from a bimodal distribution to an unimodal distribution around w_d.

Integrating back by parts the last expression we obtain the following Fokker–Planck equation

$$\frac{\partial}{\partial t} f + \frac{\partial}{\partial w}\mathcal{H}[f](w)f(w) + \frac{\partial}{\partial w}\mathcal{K}[f](w)f(w)\ dv = \frac{\varsigma}{2}\frac{\partial^2}{\partial w^2}(D(w)^2 f(w)), \tag{4.48}$$

where

$$\mathcal{K}[f](w) = \int_{\mathcal{I}} P(w,v)(v-w)f(v)\ dv, \tag{4.49}$$

$$\mathcal{H}[f](w) = \frac{4}{\kappa}\int_{\mathcal{I}}\left(w_d - \frac{w+v}{2}\right) f(v)\ dv = \frac{4}{\kappa}\left(w_d - \frac{w+m}{2}\right). \tag{4.50}$$

Remark 4.4.1 *The ratio between $\sigma^2/\alpha = \varsigma$ is of paramount importance in order to obtain in the limit the contribution of both controlled compromise propensity and diffusion [155]. Other limiting behaviors can be considered like diffusion dominated ($\varsigma \to \infty$) or controlled compromise dominated ($\varsigma \to 0$).*

4.4.2 Stationary solutions

In this section we analyze the steady solutions of the Fokker–Planck model (4.48), for particular choices of the microscopic interaction of the Boltzmann dynamic.

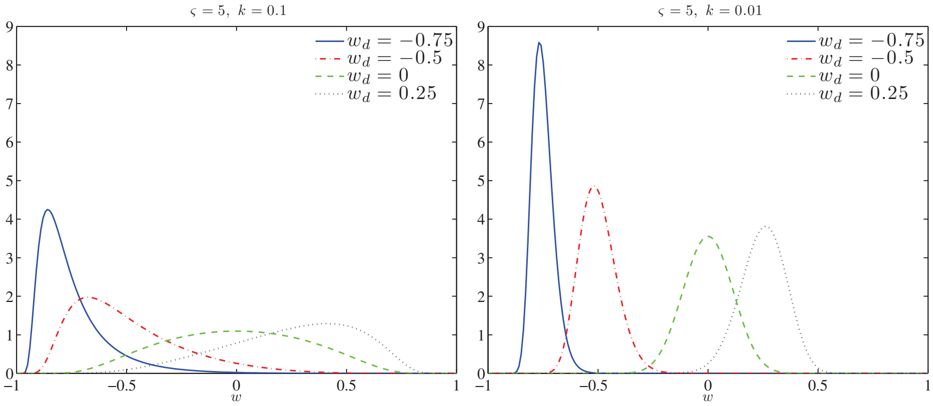


Figure 4.2: Steady state solutions in the controlled case for different values of κ and w_d. From left to right we change values of $\kappa = 0.1$ and $\kappa = 0.01$ for a fixed value of $\varsigma = 5$ and different desired states $w_d = \{-0.75, -0.5, 0, 0.25\}$.

Let consider the case in which $P(w, v) = 1$. In presence of the control the average opinion in general is not conserved in time, but since $m(t)$ converges exponentially in time to w_d, the steady state opinion solves

$$\frac{\varsigma}{2}\partial_w(D(w)^2 f) = \left(1 + \frac{2}{\kappa}\right)(w_d - w)f. \tag{4.51}$$

If we now consider as diffusion function $D(w) = (1-w^2)$, then it is possible explicitly compute the solution of (4.51) as follows [155]

$$f_\infty^\kappa(w) = \frac{C_{w_d,\varsigma,\kappa}}{(1-w^2)^2}\left(\frac{1+w}{1-w}\right)^{m/(2\varsigma)} \exp\left\{-\frac{1-mw}{\varsigma\,(1-w^2)}\left(1+\frac{2}{\kappa}\right)\right\} \tag{4.52}$$

where $C_{w_d,\varsigma,\kappa}$ is a normalization constant such that $\int f_\infty\, dw = 1$. We remark that the solution is such that $f(\pm 1) = 0$, moreover due to the general non symmetry of f, the desired state reflects on the steady state through the mean opinion. Note that in the case $\kappa \to \infty$ we obtain the steady state of the uncontrolled equation [155]. We denote by $f_\infty(w)$ this latter uncontrolled stationary behavior. We plot in Figure 4.7 the steady profile f_∞ and f_∞^κ for different choices of the parameters κ and ς. The initial average opinion $m(0)$ is taken equal to the desired opinion w_d, in this way we can see that for $\kappa \to \infty$ the constrained steady profile approaches the unconstrained one, $f_\infty^\kappa \to f_\infty$. On the other hand small values of κ give the desired distribution concentrated around w_d.

In Figure 4.8 we show the steady profile f_∞^κ for different choice of the parameters κ and the desired state w_d. We can see that decreasing the value of κ lead the profiles to concentrate around the requested value of w_d.

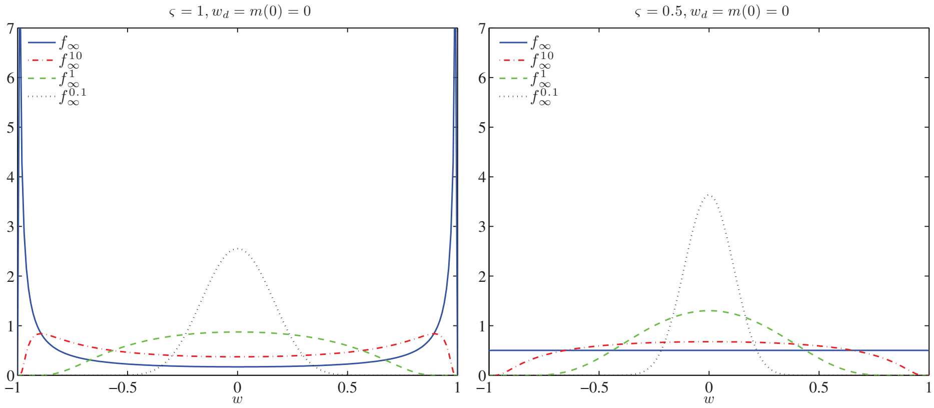


Figure 4.3: Continuous line and dashed lines represent the steady solutions f_∞ and f_∞^κ, respectively. On the left $w_d = m(0) = 0$ with diffusion parameter $\varsigma = 0.9$, on the right $w_d = m(0) = 0$ with diffusion parameter $\varsigma = 0.5$, in this last case note that f_∞ is a uniform distribution on $[-1, 1]$.

Let consider $P(w, v) = P(w)$ then stationary solutions of (4.48) satisfy the following

$$\frac{\varsigma}{2}\partial_w(D(w)^2 f) = \left(P(w) + \frac{2}{\kappa}\right)(w_d - w)f. \tag{4.53}$$

Taking $P(w) = 1 - w^2$ and $D(w) = 1 - w^2$ we can compute [155]

$$f_\infty^\kappa(w) = C_{\varsigma,m}(1-w)^{-2-\frac{w_d-1}{\varsigma}-\frac{w_d}{\kappa\varsigma}}(1+w)^{-2+\frac{w_d+1}{\varsigma}+\frac{w_d}{\kappa\varsigma}} \exp\left\{-\frac{2}{\kappa}\frac{1-w_d w}{\varsigma\,(1-w^2)}\right\} \tag{4.54}$$

We present in Figure 4.3 different profiles of f_∞^κ for $m(0) = w_d$, where we switch from the steady profile of the uncontrolled case to the steady profile (4.54).

4.5 Other constrained kinetic models

The constrained binary collision rule (4.15) admits several variants accordingly to the different ways we realize the diffusion and control dynamics.

From the modeling point of view we decided to introduce noise at the level of the explicit binary formulation (4.12),(4.14) as an external factor which can not be affected by the opinion maker. In contrast, adding noise from the very beginning in (4.1)-(4.2), or equivalently in the implicit formulation (4.5a)-(4.5b), would imply a different action of the control over the spreading of the noise. More precisely, for the binary interaction model this will originate the dynamic

$$\begin{aligned} w^* &= (1-\alpha P(w,v))\,w + \alpha P(w,v)v - \frac{\beta}{2}\left((v-w_d)+(w-w_d)\right) \\ &\quad - \alpha\frac{\beta}{2}((P(w,v)-P(v,w))(w-v)) + \left(1-\frac{\beta}{2}\right)\Theta_1 D(w) - \frac{\beta}{2}\Theta_2 D(v), \\ v^* &= (1-\alpha P(v,w))\,v + \alpha P(v,w)w - \frac{\beta}{2}\left((v-w_d)+(w-w_d)\right) \\ &\quad - \alpha\frac{\beta}{2}((P(v,w)-P(w,v))(v-w)) + \left(1-\frac{\beta}{2}\right)\Theta_2 D(v) - \frac{\beta}{2}\Theta_1 D(w). \end{aligned} \tag{4.55}$$

For this binary dynamic preservation of the bounds is more delicate and the corresponding Boltzmann model is typically written using the kernel (4.26). Note, however, that in the quasi-invariant opinion limit due to the rescaling (4.39) we have $\beta \to 0$ and therefore the limiting Fokker-Planck equation is again (4.48).

Next we remark that the microscopic constrained system (4.5a)-(4.5b) can be written in explicit form by solving the corresponding linear system for $w_1^{n+1}, \dots, w_N^{n+1}$. Straightforward computations yields the explicit formulation

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{N}\sum_{j=1}^N P_{ij}^n(w_j^n - w_i^n) + \Delta t u^n, \qquad w_i^0 = w_{0i}, \tag{4.56}$$

where now

$$u^n = \frac{(\Delta t)^2}{\nu + (\Delta t)^2}\left(\frac{1}{N^2}\sum_{h,j=1}^N P(w_h,w_j)(w_j^n - w_h^n)\right) + \frac{\Delta t}{\nu+(\Delta t)^2}(w_d - m^n), \tag{4.57}$$

and we denoted by

$$m^n = \frac{1}{N}\sum_{j=1}^N w_j^n$$

the mean opinion value. This show that a different way to realize the constrained binary dynamic (4.15) is given by

$$\begin{aligned} w^* &= (1-\alpha P(w,v))\,w + \alpha P(w,v)v - \beta\,(m(t)-w_d) \\ &\quad - \alpha\frac{\beta}{2}((P(w,v)-P(v,w))(w-v)) + \Theta_1 D(w), \\ v^* &= (1-\alpha P(v,w))\,v + \alpha P(v,w)w - \beta\,(m(t)-w_d) \\ &\quad - \alpha\frac{\beta}{2}((P(v,w)-P(w,v))(v-w)) + \Theta_2 D(w). \end{aligned} \tag{4.58}$$

Again preservation of the bounds is a difficult task and the Boltzmann equation is written in the general form (4.25). Performing the same computations as in Section

4.4.1 we obtain the limiting Fokker-Planck equation (4.48) with the simplified control term

$$\mathcal{H}[f](w) = \frac{4}{\kappa}\,(w_d - m)\,. \tag{4.59}$$

The main difference now, is that when $m(t) \to w_d$ the contribution of the control vanish, $\mathcal{H}[f](w) \to 0$, and the steady states corresponds to those of the unconstrained equation by Toscani [155] in the case where the mean opinion is given by the desired state. In other words, in the examples of Section 4.4.2, they are given by (4.52) and (4.54) in the limit case $\kappa \to \infty$. Therefore, in this case, the action of the control is weaker, since it is not able to act on any opinion distribution with mean opinion given by the desired state.

Finally, from system (4.10)-(4.11), we can also generalize (4.15) with an agent dependent action of the control. Following the same derivation as in Section 3 we have the binary interaction rule

$$\begin{aligned} w^* &= (1-\alpha P(w,v))\,w + \alpha P(w,v)v - \frac{\beta(w,v)}{2}\,(Q(v)(v-w_d) + Q(w)(w-w_d)) \\ &\quad - \alpha\frac{\beta(w,v)}{2}(Q(w)P(w,v) - Q(v)P(v,w))(v-w) + \Theta_1 D(w), \\ v^* &= (1-\alpha P(v,w))\,v + \alpha P(v,w)w - \frac{\beta(v,w)}{2}\,(Q(v)(v-w_d) + Q(w)(w-w_d)) \\ &\quad - \alpha\frac{\beta(v,w)}{2}(Q(v)P(v,w) - Q(w)P(w,v))(w-v) + \Theta_2 D(v), \end{aligned} \tag{4.60}$$

where

$$\beta(w,v) = \frac{4\alpha^2 Q(w)}{\nu + 2\alpha^2(Q(v)^2 + Q(w)^2)},$$

with property $\beta(w,v)Q(v) = \beta(v,w)Q(w)$. In this case, sufficient condition for the preservation of the bounds can be found provided that a minimal action of the control is admitted by the agents, namely assuming that $0 < Q(\,\cdot\,) \leqslant 1$. Under the scaling (4.39) we obtain the general Fokker-Plank equation (4.48) where now the control term reads

$$\mathcal{H}[f](w) = \left(\frac{2}{\kappa}\int_{\mathcal{I}} (Q(w)(w_d - w) + Q(v)(w_d - v))\,f(v)\,dv\right) Q(w). \tag{4.61}$$

4.6 Numerical examples

In this section we report some numerical test obtained by solving the constrained Boltzmann equation with the binary interaction rule (4.15) for different kind of

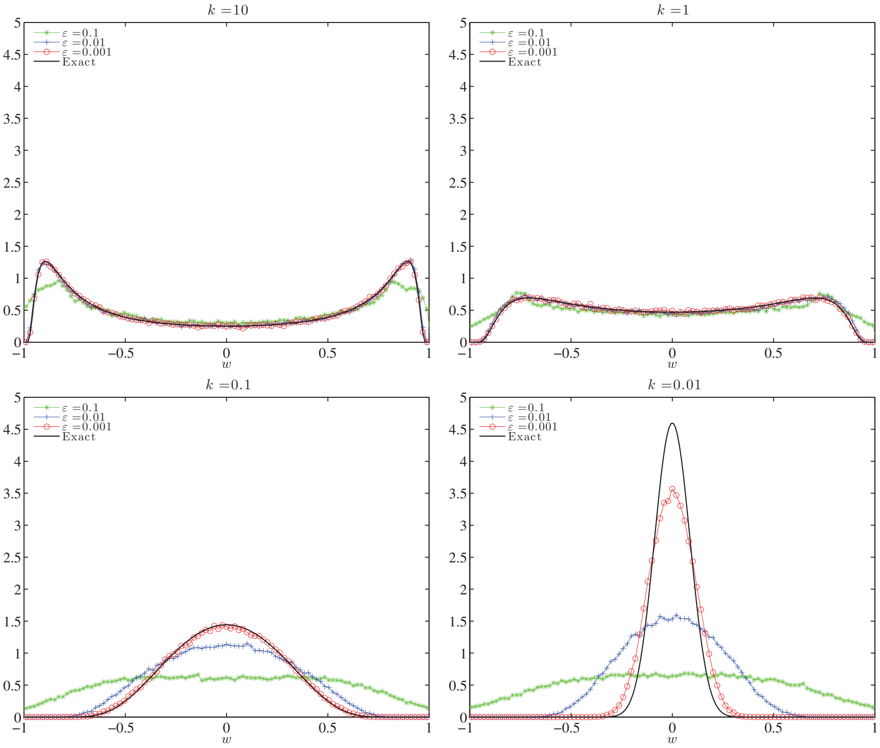


Figure 4.4: Steady solutions of the Boltzmann equation with $P(w, v) = 1$ and $D(w) = 1 - w^2$ in the scaling (4.39) for different values of ε and $\varsigma = 3$. Continuous lines represent the steady profile of the Fokker–Planck equation. From left to right from top to bottom, we increase the control action, diminishing the value of κ.

opinion models. In the numerical simulations we use a Monte Carlo methods as described in Chapter 4 of [146]. We simulate equation (4.48) for particular choices of the parameters of the model comparing the stationary solutions obtained in absence of control [155, 8] with different increasing actions of the control term.

Quasi-invariant opinion limit

In the first numerical example we compare the solutions obtained with the Monte Carlo method in the quasi-invariant opinion limit with the exact profile of the steady solution of the Fokker–Planck model (4.48). We consider the particular case

$$P(w, v) = 1, \qquad D(w) = 1 - w^2, \tag{4.62}$$

then exact solutions are described by (4.52).

In Figure 4.4 we simulate the evolution of the probability density $f(w, t)$, using a sample of $N_s = 10^5$ agents each of them interacting through the binary dynamic

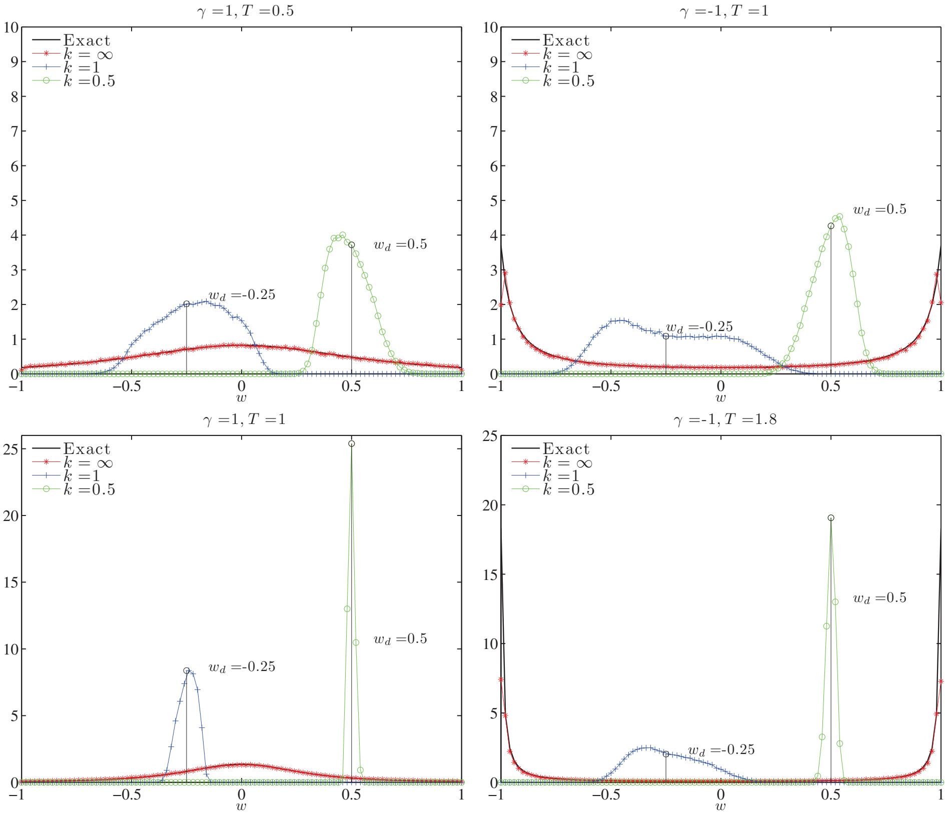


Figure 4.5: Sznajd-type model at different times. The effect of concentration ($\gamma = 1$) on the left, and separation ($\gamma = -1$) are visible for the uncontrolled case ($\kappa = \infty$). The action of a mild control $\kappa = 1$ and a strong control $\kappa = 0.1$ forces the dynamic towards different desired states, respectively $w_d = -0.25$ and $w_d = 0.5$. As expected the process needs a larger amount of time to control the separation dynamic.

(4.43) for different scaling values ε and Θ distributed uniformly on $(-\sigma, \sigma)$, with $\sigma^2 = 3\varepsilon\varsigma$, $\varsigma = 3$. Note that the discrepancy of the steady profiles in Figure 4.4 is due to the fact we are simulating the convergence of the Boltzmann equation towards its Fokker-Planck limit. Therefore decreasing ε and increasing the size of the sample N_s we can obtain better approximations of the Fokker–Planck profiles.

Sznajd-type model

In this test we consider a compromise propensity of the form

$$P(w,v) = \gamma(1 - w^2), \quad \gamma \in \mathbb{R} \tag{4.63}$$

in absence of diffusion $D(w) = 0$. Note that, when the initial mean opinion $m(0) = 0$, the quasi-invariant opinion limit in absence of control is governed by the mean-field

	$w_d = 0.25$	$w_d = 0.5$	$w_d = 0.75$	$w_d = 0.95$
$\kappa = 10$	1.7139e-01	3.428e-01	5.1351e-01	6.5032e-01
$\kappa = 5$	1.1468e-01	2.2653e-01	3.3844e-01	4.2362e-01
$\kappa = 1$	1.0592e-03	1.6027e-03	1.5460e-03	1.2877e-03
$\kappa = 0.5$	7.0990e-07	9.0454e-07	6.9543e-07	4.9742e-07

Table 4.1: L_2 distance between w_d and the average opinion m at time $T = 2$ for the controlled Sznajd-type model with separation interactions.

Sznajd's model [154, 8]

$$\partial_t f = \gamma \partial_w \left(w(1 - w^2) f \right) . \tag{4.64}$$

The model (4.64) can be solved explicitly and gives [8]

$$f(w,t) = \frac{e^{-2\gamma t}}{((1-w^2)e^{-2\gamma t} + w^2)^{3/2}} f_0 \left(\frac{w}{((1-w^2)e^{-2\gamma t} + w^2)^{1/2}} \right), \tag{4.65}$$

where $f_0(x)$ is the initial distribution. For $\gamma > 0$ we have *concentration* of the profile around zero, conversely for $\gamma < 0$ a *separation* phenomena is observed and the distribution tends to concentrate around $w = 1$ and $w = -1$.

We simulate the binary dynamic with control corresponding to the above choices starting from an initial mean opinion $m(0) = 0$. Our aim is to explore the differences between the controlled concentration and separation dynamics. We choose a scaling parameter $\varepsilon = 0.005$ and a number of sample agents of $N = 10^5$.

In Figure 4.5 we simulate the evolution of $f(w,t)$ for the concentration ($\gamma = 1$) and separation ($\gamma = -1$) cases. Starting from the uniform distribution on $\mathcal{I}$, we investigate three different cases: uncontrolled ($\kappa = \infty$), mild control ($\kappa = 1$) towards desired state $w_d = -0.25$ and strong control ($\kappa = 0.1$) towards $w_d = 0.5$. The solution profiles in the uncontrolled case, $\kappa = \infty$ coincides with the exact solution profile given by (4.65). Observe that separation phenomena implies a slower convergence towards the desired states.

We complete the tests just presented with Table 4.1, where we measure the L^2 distance between the average opinion m at final time $T = 2$ and the desired state w_d, in the separation case, ($\gamma = -1$). We compare the errors for decreasing values of κ and for different values of the desired state w_d, showing that more effective control implies faster convergence.

Bounded confidence model

Next, we consider the case of *bounded confidence models*, where the possible interaction between agents depends on the level of confidence they have [102, 97]. This can be model through a compromise function which accounts the exchange of opinion only inside a fixed distance Δ between the agent opinions

$$P(w,v) = \chi(|w-v| \leqslant \Delta), \tag{4.66}$$

where $\chi(\,\cdot\,)$ is the indicator function.

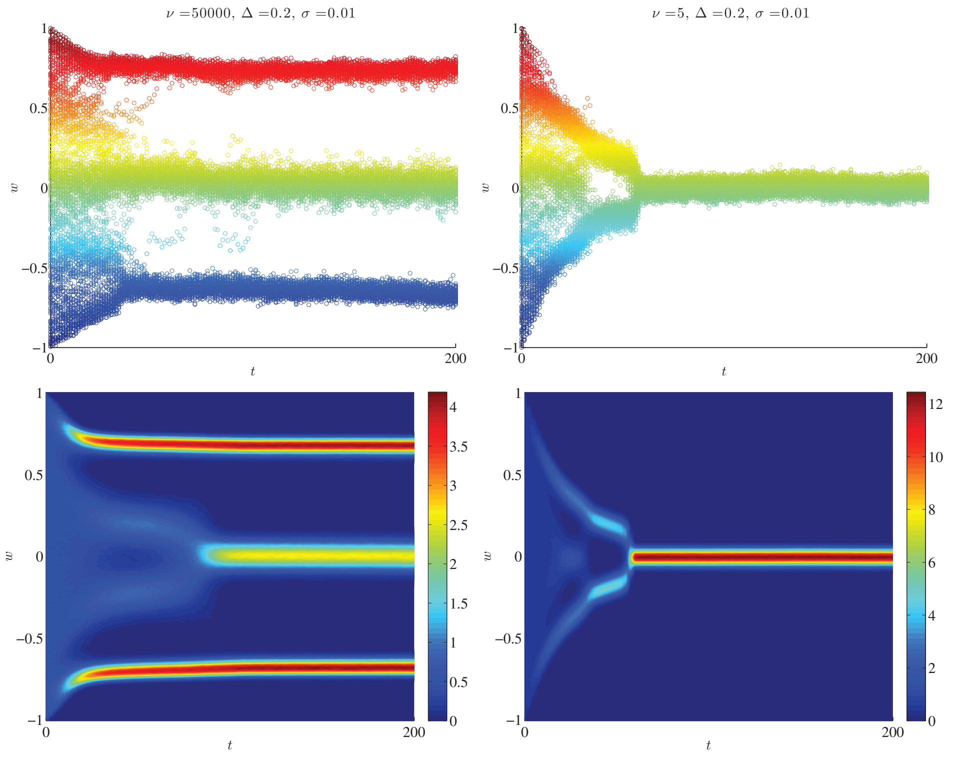


Figure 4.6: Bounded confidence model. On the left the control parameter $\nu = 5000$ on the right $\nu = 5$. In the top row the result of a particle simulation with $N = 200$ agents where the color scale depicts the opinion value. Bottom row represents the evolution of the kinetic density. In both cases the simulation is performed for $\sigma = 0.01$ and $\Delta = 0.2$.

In Figure 4.6, we simulate the dynamic of the agents starting from an uniform distribution of the opinions on the interval $\mathcal{I} = [-1, 1]$. The confidence bound is taken $\Delta = 0.2$ and the diffusion parameter $\sigma = 0.01$. We consider the case without control and with control, letting the system evolve in the time interval $[0\ T]$, with $T = 200$. In the left column figures we represents the weak controlled case, with penalization parameter $\nu = 5000$, and three mainstream opinions emerge, on the right the presence of the control, $\nu = 5$ is able to lead the opinions to concentrate around the desired opinion, $w_d = 0$.

Top row of plots shows the evolution of the dynamic at the particle level, with $N = 200$. Bottom row represents the same dynamic at the kinetic level, simulation is performed with a sample of $N_s = 2 \times 10^5$ particles with $\varepsilon = 0.05$.

4.7 Mean-field model predictive control of flocking behavior

In the following section we extend the results of the previous sections to the case of second-order models for alignment, [158, 67, 138].

We consider the evolution of N agents where each agent has position and velocity $(x_i(t), v_i(t))$, $i = 1, \ldots, N$ and this can change over time according to

$$\begin{aligned} &\dot{x}_i = v_i, \\ &\dot{v}_i = \frac{1}{N}\sum_{j=1}^{N} H_a(x_i, x_j)(v_j - v_i) + u, \qquad\qquad v_i(0) = v_{0i}, \end{aligned} \tag{4.67}$$

where at variance with respect to Cucker-Smale model, function H_a is such that $H_a(x, y) \neq H_a(x, y)$. As before the control $u = u(t)$ is given by the minimization of the cost functional over a certain time horizon T

$$u = \text{argmin} \int_0^T \frac{1}{N}\sum_{j=1}^{N} \left(\frac{1}{2}(v_j(s) - v_d)^2 + \frac{\nu}{2}(u(s))^2 \right) ds, \tag{4.68}$$

In the formulation (4.2) $\nu > 0$ is a regularization parameter and the value v_d is a desired velocity, which can be also extended to more general objects, for example to a time dependent velocity $v_d = v_d(t)$ or the direction along some desired trajectory.

4.7.1 MPC for flocking models

We use instantaneous control to derive a computable control u at any time t, which results to be suboptimal respect to the solution to (4.67)-(4.68).

A receding horizon strategy

Following the same approach of section 4.2.1, we split the time interval $[0, T]$ in M time intervals of length Δt and let $t^n = \Delta t\ n$, assuming the control piecewise constant on time intervals of length $\Delta t > 0$,

$$u(t) = \sum_{n=0}^{M-1} u^n \chi_{[t^n, t^{n+1}]}(t).$$

We determine the value of the control $u^n \in \mathbb{R}$, solving for a state $\bar{v}_i$ the (reduced) optimization problem

$$\begin{aligned} \dot{v}_i &= \frac{1}{N}\sum_{j=1}^{N} H_a(x_i,x_j)(v_j - v_i) + u, && v_i(t^n) = \bar{v}_i, \\ u^n &= \mathrm{argmin}_{u\in\mathbb{R}} \int_{t^n}^{t^{n+1}} \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{2}(v_j - v_d)^2 + \frac{\nu}{2}u^2\right) ds, && u \in [u_L, u_R]. \end{aligned} \tag{4.69}$$

Having the control u^n on the interval $[t^n, t^{n+1}]$, we let evolve v_i according to the dynamics

$$\dot{v}_i = \frac{1}{N}\sum_{j=1}^{N} H_a(x_i,x_j)(v_j - v_i) + u^n \tag{4.70}$$

in order to obtain the new state $\bar{v}_i = v_i(t^{n+1})$. We again solve (4.69) to obtain u^{n+1} with the modified initial data and we repeat this procedure until we reach $n\Delta t = T$.

In this way we are able to reduce the complexity of the initial problem (4.67)-(4.68), to an optimization problem in a single real–valued variable u^n. Moreover the quadratic cost and a suitable discretization of (4.70) allows an explicit representation of u^n in terms of $\bar{v}_i$ and $v_i(t^{n+1})$. As shown in section 4.2.2 this allows to reformulate the previous algorithm as a feedback controlled system which in discretized form reads

$$\begin{aligned} v_i^{n+1} &= v_i^n + \frac{\Delta t}{N}\sum_{j=1}^{N} H_{a,ij}^n (v_j^n - v_i^n) + \Delta t u^n, && v_i^n = \bar{v}_i, \\ u^n &= -\frac{\Delta t}{\nu N}\sum_{j=1}^{N}(v_j^{n+1} - v_d). \end{aligned} \tag{4.71}$$

where $H_{a,ij}^n = H_a(x_i^n, x_j^n)$. Therefore the feedback controlled system in the discretized form results

$$\begin{aligned} x_i^{n+1} &= x^n + \Delta t v^n \\ v_i^{n+1} &= v_i^n + \frac{\Delta t}{N}\sum_{j=1}^{N} H_{a,ij}^n (v_j^n - v_i^n) - \frac{(\Delta t)^2}{\nu N}\sum_{j=1}^{N}(v_j^{n+1} - v_d), && v_i^n = \bar{v}_i. \end{aligned} \tag{4.72}$$

Where the action of the control is substituted by a relaxation term toward the desired velocity v_d, appearing in an implicit form.

4.7.2 Mean-field description

The aim of this section is to show that out of system (4.72) we are able to derive a corresponding kinetic approximation of the feedback controlled system.

Derivation of the forward system

Let us introduce the parameter $\gamma = \Delta t/N$, we can write the second part of system (4.72) as follows

$$v_i^{n+1} = v_i^n + \gamma \sum_{j=1}^{N} H^n_{a,ij}(v_j^n - v_i^n) - \frac{N\gamma^2}{\nu} \sum_{j=1}^{N}(v_j^{n+1} - v_d), \qquad v_i^0 = v_{0i}, \tag{4.73}$$

in matrix-vector notation we have

$$\left(\text{Id} + \frac{\gamma^2}{\nu}\mathbf{E}\right) v^{n+1} = v^n + \gamma \mathbf{H}_a^n v^n - \gamma \mathbf{D}^n v^n - \frac{N\gamma^2}{\nu}\mathbf{e} v_d \tag{4.74}$$

where

$$\mathbf{H}_a^n = \begin{pmatrix} H^n_{a,11} & H^n_{a,12} & \dots & H^n_{a,1N} \\ \vdots & & & \vdots \\ H^n_{a,N1} & H^n_{a,N2} & \dots & H^n_{a,NN} \end{pmatrix}, \quad \mathbf{D}^n = \begin{pmatrix} \sum_j H^n_{a,1j} & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & \sum_j H^n_{a,Nj} \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \vdots & & & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}, \quad \mathbf{e} = (1, 1, \dots, 1)^T, \quad v^n = (v_1^n, v_2^n, \dots, v_N^n)^T.$$

The system can be reverted in a fully explicit form

$$v^{n+1} = \mathbf{A} v^n + \gamma \mathbf{A} \mathbf{H}_a^n v^n - \gamma \mathbf{A} \mathbf{D}^n v^n - \frac{N\gamma^2}{\nu} \mathbf{A} \mathbf{e} v_d \tag{4.75}$$

where $\mathbf{A}$ has the following structure and property

$$\mathbf{A} = \left(\text{Id} + \frac{\gamma^2}{\nu}\mathbf{E}\right)^{-1} = \text{Id} - \frac{\gamma^2}{\nu + N\gamma^2}\mathbf{E}, \qquad \mathbf{A}\mathbf{e} = \frac{\nu}{\nu + N\gamma^2}\mathbf{e}.$$

therefore

$$\begin{aligned} \mathbf{A}\mathbf{H}^n &= \mathbf{H}_a^n - \frac{\gamma^2}{\nu + N\gamma^2}\mathbf{E}\mathbf{H}_a^n = \mathbf{H}_a^n - \frac{\gamma^2}{\nu + N\gamma^2}\bar{\mathbf{H}}_a^n, \\ \mathbf{A}\mathbf{D}^n &= \mathbf{D}^n - \frac{\gamma^2}{\nu + N\gamma^2}\mathbf{E}\mathbf{D}^n = \mathbf{D}^n - \frac{\gamma^2}{\nu + N\gamma^2}(\bar{\mathbf{H}}_a^n)^T, \end{aligned}$$

where we indicate $\bar{\mathbf{H}}_a^n$ the matrix product $\mathbf{E}\mathbf{H}_a^n = (\mathbf{E}\mathbf{D}^n)^T$. The full vector-system reads

$$\begin{aligned} v^{n+1} = v^n + \gamma\left(\mathbf{H}_a^n - \mathbf{D}^n\right) v^n + \frac{N\gamma^2}{\nu + N\gamma^2}\, v_d \mathbf{e} - \frac{\gamma^2}{\nu + N\gamma^2}\mathbf{E} v^n + \\ - \frac{\gamma^3}{\nu + N\gamma^2}(\bar{\mathbf{H}}_a^n - (\bar{\mathbf{H}}_a^n)^T) v^n, \end{aligned} \tag{4.76}$$

where the last term disappears if function $H_a = H_a(x,y)$ is symmetric. Reverting back In terms of Δt the full system reads have

$$\begin{aligned} x_i^{n+1} &= x^n + \Delta t v^n, \\ v_i^{n+1} &= v_i^n + \frac{\Delta t}{N}\sum_{j=1}^N H^n_{a,ij}(v_j^n - v_i^n) + \frac{\Delta t^2}{\nu + \Delta t^2}\frac{1}{N}\sum_{j=1}^N \left(v_d - v_j^n\right) + \\ &\quad - \frac{\Delta t^3}{\nu+\Delta t^2}\frac{1}{N^2}\sum_{j=1}^N (H^n_{a,ji} - H^n_{a,ij})v_i^n. \end{aligned} \tag{4.77}$$

For $\Delta t \to 0$ this corresponds to a time discretization of the original system (4.69), where the action of the control is lost since it is expressed in terms of $O(\Delta t^2)$.

In order to see the control action is necessary to assume the following scaling on the regularization parameter, $\nu = \Delta t \kappa$, thus we obtain

$$v_i^{n+1} = v_i^n + \frac{\Delta t}{N}\sum_{j=1}^N H^n_{a,ij}(v_j^n - v_i^n) + \frac{\Delta t}{\kappa + \Delta t}\frac{1}{N}\sum_{j=1}^N \left(v_d - v_j^n\right) + O(\Delta t), \tag{4.78}$$

Letting $\Delta t \to 0$ we obtain the full controlled continuous system as follows

$$\begin{aligned} \dot{x}_i &= v_i \\ \dot{v}_i &= \frac{1}{N}\sum_{j=1}^N H_a(x_i,x_j)(v_j - v_i) + \frac{1}{\kappa N}\sum_{j=1}^N (v_d - v_j)\,. \end{aligned} \tag{4.79}$$

At this level, in order to derive a corresponding kinetic approximation, as done in section 4.3 we could proceed approximating the dynamic through a binary interaction and deriving the corresponding Boltzmann-Povzner equation. At variance with this approach we show in the following a direct approximation through the *mean-field limit*, [47].

4.7.3 Mean-field model predictive control limit

We want to give a kinetic description of (4.79), therefore we introduce function $f = f(x,v,t)$, representing the particle density at time t with position and velocity (x,v). We assume that f is equivalent to a probability density function on $\mathbb{R}^{2d}$, thus

$$\int_{\mathbb{R}^{2d}} f(x,v,t)\; dxdv = 1.$$

Since the right term of system (4.79) is assumed to be continuos, bounded and locally Lipschitz, it satisfies all the nice properties of classical swarming models, see [47]. Therefore the derivation of a *mean-field* equation follows from straight forward computations.

Mean-field limit derivation

We define the empirical measures

$$f^N(t) = \frac{1}{N}\sum_{i=1}^N \delta(x - x_i(t))\delta(v - v_i(t)),$$

and let consider a test function $\phi \in \mathcal{C}_0^1(\mathbb{R}^{2d})$ and we compute

$$\begin{aligned}\frac{d}{dt}\left\langle f^N(t), \phi\right\rangle = \frac{1}{N}\sum_{i=1}^N \frac{d}{dt}\phi(x_i(t), v_i(t)) = \\ \underbrace{\frac{1}{N}\sum_{i=1}^N \nabla_x\phi(x_i(t), v_i(t)) \cdot v_i(t)}_{I_1} + \underbrace{\frac{1}{N^2}\sum_{i,j=1}^N \nabla_v\phi(x_i(t), v_i(t))H_a(x_i, x_j)(v_j - v_i)}_{I_2} + \\ + \underbrace{\frac{1}{\kappa N^2}\sum_{i,j=1}^N \nabla_v\phi(x_i(t), v_i(t))(v_d - v_i)}_{I_3}.\end{aligned}$$

We solve term by term the summation, thus

$$\begin{aligned} I_1 &= \frac{1}{N}\sum_{i=1}^N \nabla_x\phi(x_i(t), v_i(t)) \cdot v_i(t) = \left\langle f^N(t), \nabla_x\phi \cdot v\right\rangle, \\ I_2 &= \frac{1}{N^2}\sum_{i,j=1}^N H_a(x_i(t), x_j(t))\nabla_v\phi(x_i(t), v_i(t)) \cdot (v_j(t) - v_i(t)) = \\ &\qquad \left\langle f^N(t), \frac{1}{N}\sum_{j=1}^N H(|x - x_j(t)|)\nabla_v\phi \cdot (v_j(t) - v) \right\rangle. \end{aligned}$$

and

$$I_3 = \frac{1}{\kappa N^2}\sum_{i,j=1}^N \nabla_v\phi(x_i(t), v_i(t)) \cdot (v_d - v_i(t)) = \left\langle f^N(t), \frac{1}{\kappa N}\sum_{j=1}^N \nabla_v\phi \cdot (v_d - v) \right\rangle.$$

Defining the following quantities

$$\begin{aligned} \rho^N(x, t) &:= \int_{\mathbb{R}^d} f^N(x, v, t)dv = \frac{1}{N}\sum_{i=1}^N \delta(x - x_i(t)), \\ m^N(x, t) &:= \int_{\mathbb{R}^d} v f^N(x, v, t)dv = \frac{1}{N}\sum_{i=1}^N v_i(t)\delta(x - x_i(t)), \end{aligned}$$

we can express the previous relation in the following way

$$I_2 = \left\langle f^N(t), \nabla_v \phi \cdot H \star m^N \right\rangle - \left\langle f^N(t), \nabla_v \phi \cdot H \star \rho^N \right\rangle$$
$$I_3 = \left\langle f^N(t), \nabla_v \phi \cdot \frac{1}{\kappa}(v_d - m^N) \right\rangle$$

where $\star$ represents the convolution product. Collecting all the terms and integrating by parts in (x, v) we recover the following weak formulation

$$\frac{d}{dt}\left\langle f^N, \phi \right\rangle = -\left\langle v\nabla_x f^N, \phi \right\rangle - \left\langle \nabla_v \mathcal{H}[f^N]f^N, \phi \right\rangle - \left\langle \nabla_v \mathcal{K}[f^N]f^N, \phi \right\rangle$$

where

$$\mathcal{H}[f^N](x, v, t) = \int_{\mathbb{R}^{2d}} H_a(x, y)(v - w)f^N(y, w, t)dydw,$$
$$\mathcal{K}[f^N](v, t) = \frac{1}{\kappa}\int_{\mathbb{R}^{2d}} (w - v_d)f^N(y, w, t)dydw.$$

Rewriting the main expression we have

$$\left\langle \frac{\partial}{\partial t} f^N + v\nabla_x f^N + \nabla_v \mathcal{H}[f^N]f^N + \nabla_v \mathcal{K}[f^N]f^N, \phi \right\rangle = 0$$

and thus the strong form reads

$$\frac{\partial}{\partial t} f^N + v\nabla_x f^N + \nabla_v \mathcal{H}[f^N]f^N + \nabla_v \mathcal{K}[f^N]f^N = 0.$$

Then the limit for $k \to \infty$ of subsequence $(f^{N_k})_k$ leads to the following evolution equation for $f = f(x, v, t)$

$$\begin{aligned} \partial_t f + v \cdot \nabla_x f &= -\nabla_v \cdot (\mathcal{H}[f]f) - \nabla_v \cdot (\mathcal{K}[f]f)\,, \\ \mathcal{H}[f] &= \int_{\mathbb{R}^{2d}} H_a(x, y)(w - v)f(y, w, t)\; dydw. \\ \mathcal{K}[f] &= \frac{1}{\kappa}\int_{\mathbb{R}^{2d}} (v_d - w)f(y, w, t)\; dydw = \frac{1}{\kappa}\left(v_d - m(t)\right). \end{aligned} \tag{4.80}$$

4.7.4 Numerical tests

In the following section we perform some numerical test using the algorithms developed in [6].

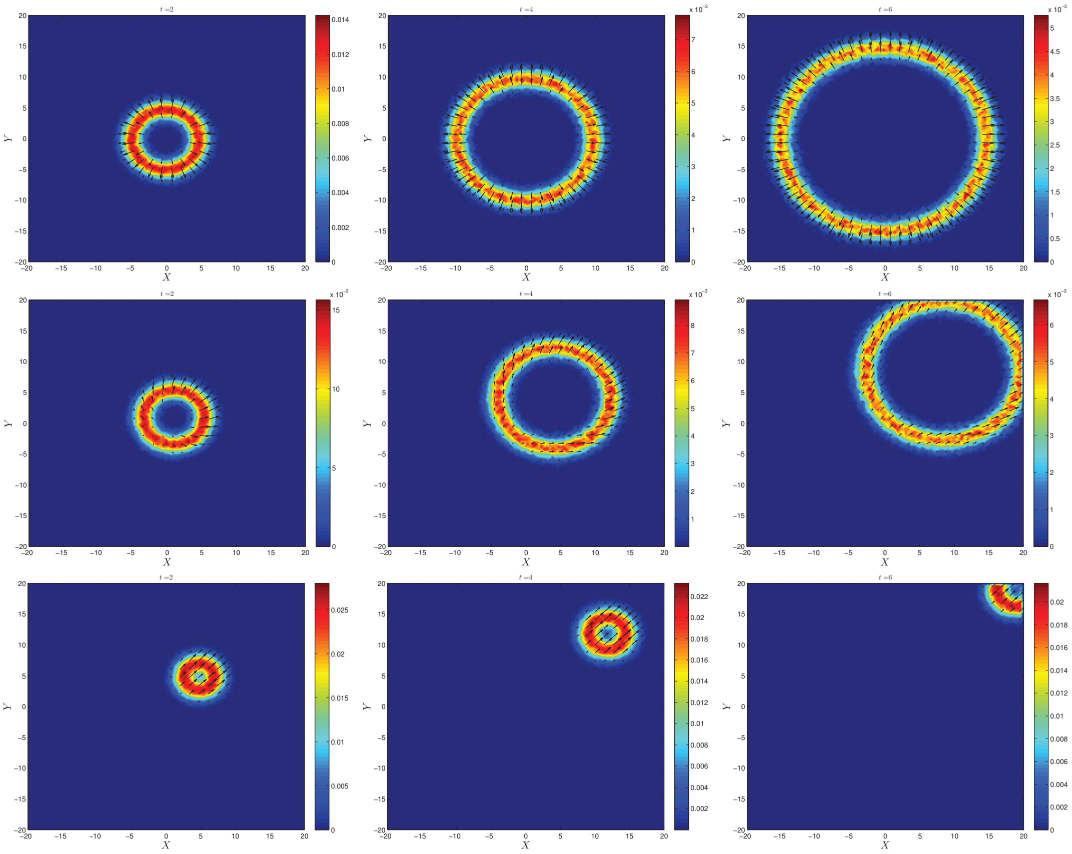


Figure 4.7: Evolution of the kinetic Cucker-Smale model with $\gamma = 10$, with different actions of the control, at three different steps. First row, $\kappa = \infty$, the alignment condition is not reached. Second row, $\kappa = 10$, (mild control action) alignment is reached towards the desired velocity $v_d = (1,1)^T$. Bottom row, strong control action, $\kappa = 1$ the control is reached quickly and the spatial density is more concentrated.

Forcing consensus

We recall that the classical alignment dynamic in Cucker-Smale model is weighted by the following not increasing function

$$H_a(x, y) = \frac{K}{(\varsigma^2 + |x - y|^2)^\gamma},$$

it has been shown that for $\gamma \leqslant 1/2$ we have *unconditional focking*, i.e. all agents tend to move exponentially fast with the same velocity, while their relative distances tend to remain constant, [54, 67]. We want to address our attention to the formation of alignment, when this condition is not satisfied by the parameter γ or by the initial conditions.

In Figure 4.7 we solve model (4.80), with $\gamma = 10$ and with desired speed $v_d = (1,1)^T$, we compare the evolution of the same initial data for different choices of control strength . The initial data, normally distributed in space and with symmetric bimodal distribution in velocity

$$f_0(x,v) = C \exp\left\{\frac{-(x^2+y^2)}{2\sigma_x}\right\}\left(\exp\left\{\frac{-(v+v_0)^2}{2\sigma_v}\right\} + \exp\left\{\frac{-(v-v_0)^2}{2\sigma_v}\right\}\right),$$

where $v_0 = 5$ and $\sigma_x = \sigma_v = 1$ and C a normalization constant. Top row shows the evolutions without control action, $\kappa = \infty$, so the density follows the velocity flux spreading in the space with a fat ring shape, without reaching a condition of alignment. Second row shows the evolution with mild control action ($\kappa = 10$) and bottom row the evolution of a strong control ($\kappa = 1$), in both cases the alignment to the desired speed is obtained, but in different time scale and spread of density.

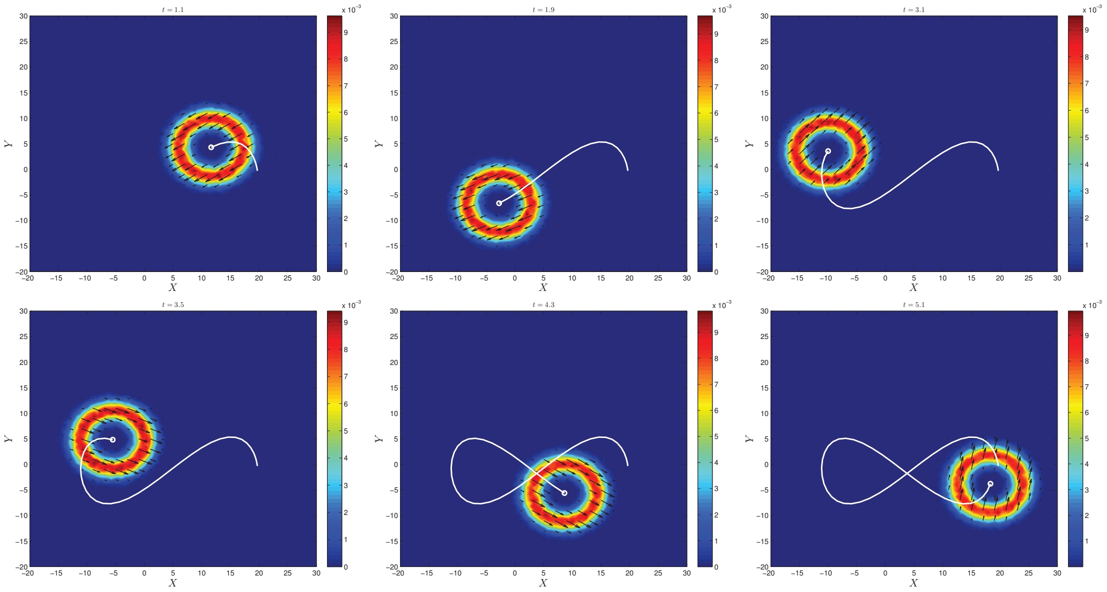


Figure 4.8: The flock density is forced to follow a desired trajectory $\gamma_d(t) = (cos(t), sin(2t))$, described by a *lemniscate*, the regularization parameter is $\kappa = 0.1$ and the scaling parameter $\varepsilon = 0.01$.

Following a desired trajectory

We can extend the forcing consensus problem to a problem of following a desired trajectory $\gamma_d(t)$, considering a desired speed as a function of time, $v_d = v_d(t)$, therefore $v_d(t) = \gamma_d'(t)$. In this way the control action at time t forces the system to converge instantaneously to the corresponding desired velocity at time t. Note that in this case the choice of κ and Δt are of paramount importance to reconstruct exactly the trajectory.

We simulate the evolution of an aligned density solution, forcing the system to follow a desired trajectory $\gamma_d(t) = R(\cos(t), \sin(2t))$ with $R = 1$, which corresponds to a *lemniscate*. The corresponding desired speed is given by $v_d(t) = R(-\sin(t), 2\cos(2t))$, which we discretize according to our algorithm with $\Delta t = 0.01$.

In Figure 4.8 we show the evolution of the system with control action, $\kappa = 0.1$.

4.8 Conclusions

In this chapter we introduced a general way to construct a Boltzmann description of optimal control problems for large systems of interacting agents. The approach has been applied to a constrained microscopic model of opinion formation. The main feature of the method is that, thanks to a model predictive approximation, the control is explicitly embedded in the resulting binary interaction dynamic. In particular in the so-called quasi invariant opinion limit simplified Fokker-Planck models have been derived which admit explicit computations of the steady states. The robustness of the controlled dynamics has been illustrated by several numerical examples which confirm the theoretical results. Different generalizations of the presented approach are possible, like the introduction of the same control dynamic through leaders or the application of this same control methodology to swarming and flocking models.

CHAPTER 5

Asymptotic Preserving schemes for the time discretization of optimal control for hyperbolic problems with relaxation

5.1 Introduction

We are interested in numerical methods for time discretization of optimal control problems of type (5.1). The construction of such methods for control problems involving differential equations has been an intensive field of research recently [36, 79, 101, 123, 162]. Applications of such methods can be found in several disciplines, form aerospace and mechanical engineering to the life sciences. In particular, many applications involves systems of differential equations of the form

$$y'(t) = f(y(t), t) + \frac{1}{\varepsilon} g(y(t), t), \tag{5.1}$$

where f and g, eventually obtained as suitable finite-difference or finite-element approximations of spatial derivatives, induce considerably different time scales indicated by the small parameter $\varepsilon > 0$ in the previous equation. Therefore, to avoid fully implicit integrators, it is highly desirable to have a combination of implicit and explicit (IMEX) discretization terms to resolve stiff and non–stiff dynamics accordingly. For Runge–Kutta methods such schemes have been studied in [12, 38, 78, 112, 115, 145, 144].

Control problems with respect to IMEX methods have been investigated also in [111, 107] in the case of fixed positive value of $\varepsilon > 0$. Among the most relevant examples for IMEX scheme are the time discretization of hyperbolic balance laws and kinetic equations. As discussed in [115, 144] the construction of such methods imply new difficulties due to the appearance of coupled order conditions and to the possible loss of accuracy close to stiff regimes $\varepsilon \ll \Delta t$ and Δt being the time discretization of the numerical scheme. In contrary to the existing work [111, 107] we focus here on optimal control problems where the time integration schemes also allow a accurate resolution in the stiff regime. As a prototype example including already the major difficulties for such methods we choose the Goldstein-Taylor model (5.3). This equation already contains several ingredients typical to linear kinetic transport models and serves as a prototype and test case for numerical integration schemes. The model describes the time evolution of two particle densities $f^+(x,t)$ and $f^-(x,t)$, with $x \in \Omega \subset \mathbb{R}$ and $t \in \mathbb{R}^+$, where $f^+(x,t)$ (respectively $f^-(x,t)$) denotes the density of particles at time $t > 0$ traveling along a straight line with velocity $+c$ (respectively $-c$). The particle may change with rate σ the direction. The differential model can be written as

$$\begin{aligned} f_t^+ + cf_x^+ &= \sigma\left(f^- - f^+\right), \\ f_t^- - cf_x^- &= \sigma\left(f^+ - f^-\right). \end{aligned} \tag{5.2}$$

Introducing the macroscopic variables

$$\rho = f^+ + f^-, \qquad j = c(f^+ - f^-)$$

we obtain the equivalent form

$$\begin{aligned} \rho_t + j_x &= 0, \\ j_t + c^2\rho_x &= -2\sigma j. \end{aligned} \tag{5.3}$$

We introduce a linear quadratic optimal control problem subject to a relaxed hyperbolic system of balance laws. Let $\Omega = [0,1]$, terminal time $T > 0$, regularisation parameter $\nu \geqslant 0$ and let $u(t)$ be the control. The function $\rho_d(x)$ is a desired state. To simplify notation we set $c^2 = 2\sigma = 1/\varepsilon^2$ and $\varepsilon > 0$ is the non–negative relaxation parameter.

The optimization problem then reads

$$\min J(\rho, u) = \frac{1}{2}\int_0^1 (\rho(x,T) - \rho_d(x))^2 dx + \frac{\nu}{2}\int_0^T u^2(t)dt \tag{5.4}$$

subject to

$$\rho_t + j_x = 0, \tag{5.5a}$$

$$j_t + \frac{1}{\varepsilon^2}\rho_x = -\frac{1}{\varepsilon^2}j. \tag{5.5b}$$

$$\rho(x,0) = \rho_0, \qquad j(x,0) = j_0 \tag{5.5c}$$

$$j(0,t) = 0, \qquad j(1,t) - \rho(1,t) = -u(t) \tag{5.5d}$$

Further, we set box constraints for the control

$$u_l(t) \leqslant u(t) \leqslant u_r(t)$$

In the limit case $\varepsilon \to 0$, (5.5b) formally yields

$$j(x,t) = -\rho_x(x,t).$$

Plugging this into (5.5a) yields the heat equation

$$\rho_t = \rho_{xx}$$

and the optimal control problem (5.4) – (5.5) reduces to a problem studied for example in [156]. Obviously, we expect a similar behavior for a numerical discretization. This property, called asymptotic preserving, has been investigated for the simulation of Goldstein–Taylor like models in [38, 78, 112] but has not yet been studied in the context of control problems.
The paper is organized as follows. In Section 5.2 we introduce the temporal discretization of problem (5.5) and describe in detail the resulting semi–discretized optimal control problems. We investigate which numerical integration schemes yield a stable approximation to the resulting optimality conditions. In the third section we show how to provide a stable discretization scheme in the parabolic regime by introducing a splitting and applying the formal Chapman-Enskog type limiting procedure. In Section 5.4 we present numerical results on the several implicit explicit Runge–Kutta methods (IMEX) schemes for the limiting problem as well as on an example taken from [156]. Definitions for properties of the IMEX schemes are collected for convenience in the appendix 5.5.

5.2 The semi–discretized problem

We are interested to derive a numerical time integration scheme which allows to treat the optimal control problem (5.4)–(5.5) for all values of $\varepsilon \in [0,1]$, including in particular the limit case $\varepsilon = 0$. Therefore, we leave a side the treatment of the discretization of the spatial variable x as well as theoretical aspects of the differentiability of solutions (ρ, J) of equation (5.5). We remark that the semigroup generated by a nonlinear hyperbolic conservation/balance law is generically non-differentiable in L^1 even in the scalar one-dimensional (1-D) case. More details on the differential structure of solutions are found in [41, 42], on convergence results for first–order numerical schemes and scalar conservation laws are found in [31, 57**?** , 94, 157] Numerical methods for the optimal control problems of *scalar* hyperbolic equations have been discussed in [20, 92, 93, 157]. In [95, 96], the adjoint equation has been discretized using a Lax-Friedrichs-type scheme, obtained by including conditions along shocks and modifying the Lax-Friedrichs numerical viscosity. Convergence of the modified Lax-Friedrichs scheme has been rigorously proved in the case of a smooth convex flux function. Convergence results have also been obtained in [157] for the class of schemes satisfying the one-sided Lipschitz condition (OSLC) and in [20] for a first–order implicit-explicit finite-volume method. To the best of our knowledge there does not exists a convergence theory for spatial discretization of control problems subject to hyperbolic systems with source terms so far.

In view of the previous discussion the interest is on the availability of suitable time–integration schemes for the arising optimal control problem. We consider therefore a semi–discretized problem in time. We further skip the spatial dependence whenever the intention is clear. The system (5.5a) consists of a stiff and a non–stiff part we employ diagonal implicit explicit Runge–Kutta methods (IMEX). Convergence order of such schemes for positive ε and the property of symplecticity has been analysed in [107]. In the following we briefly review IMEX methods and discuss a splitting [38] in order to also resolve efficiently the stiff limiting problem ($\varepsilon = 0$).

An $s-$stage IMEX Runge–Kutta method is characterized by the $s \times s$ matrices $\tilde{A}, A$ and vectors $\tilde{c}, c, \tilde{b}, b \in \mathbb{R}^s$, represented by the double Butcher tableau:

$$\text{Explicit:} \qquad \begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline & \tilde{b}^T \end{array} \qquad\qquad \text{Implicit:} \qquad \begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

We refer to the appendix 5.5 for further definitions and examples of IMEX RK schemes. Applying an IMEX time–discretization to the Goldstein-Taylor model (5.5) yields in the limit $\varepsilon = 0$ an explicit numerical scheme for the heat equation [38]. This is only stable provided the parabolic CFL condition $\Delta t \approx \Delta x^2$ holds true. This is highly undesirable and therefore, a splitting has been introduced such that also in the limit $\varepsilon = 0$ an implicit discretization of the heat equation can be obtained. We rewrite (5.5a) as

$$\rho_t = -\overbrace{(j + \mu \rho_x)_x}^{\textit{explicit}} + \overbrace{(\mu \rho_{xx})}^{\textit{implicit}} \tag{5.6}$$

where $\mu = \mu(\varepsilon) \geqslant 0$ is such that $\mu(0) = 1$ and leave equation (5.5b) unchanged. Within an IMEX time discretization we treat explicitly the first term and implicitly the second term as indicated in (5.6). It remains to discuss the choice of μ in equation 5.6 depending on ε. Using formal Chapman–Enkog expansion for this choice, presented in section 5.6, we observe that in the diffusive limit $\varepsilon = 0$ the term $j + \mu\rho_x$ vanishes.

Combining the previous computations we state the semi–discretized problem for an $s-$stage IMEX scheme. Introduce a temporal grid of size Δt and N equally spaced grid points t_n such that $T = \Delta t N$ and $t_1 = 0$. Let $\rho^n = \rho(t_n, \cdot), j^n = j(t_n, \cdot)$, $\mathbf{e} = (1, \dots, 1) \in \mathbb{R}^s$ and denote by $\mathbf{R} = (R_\ell(\cdot))_{\ell=1}^s$ the s stage variables and similarly for $\mathbf{J}$. For notational simplicity we discretize the control on the same temporal grid $u^n = u(t_n)$. However, this is not necessary for the derived results and other approaches can be used. We prescribe boundary conditions in the case $\varepsilon > 0$ as follows: Since in the limit $\varepsilon = 0$ we obtain $j(t,x) = -\rho_x(t,x)$ we add $j^n(1) = -\rho_x^n(1)$ and $j^n(0) = -\rho_x^n(0)$ as boundary conditions. Further let $\mathcal{M} = \operatorname{diag}(\mu_l) \in \mathbb{R}^{s\times s}$ define the values of μ_l for the levels $l = 1,\dots,s$.

Then, the semi–discretization of problem (5.5) reads

$$\begin{aligned}
&\min \frac{1}{2}\int_0^1 (\rho^N(x) - \rho_d(x))^2 dx + \Delta t \frac{\nu}{2}\sum_{n=1}^N (u^n)^2 \,,\\
\mathbf{R} &= \rho^n \mathbf{e} - \Delta t \tilde{A}(\partial_x \mathbf{J} + \mathcal{M}\partial_{xx}^2 \mathbf{R}) + \Delta t A\left(\mathcal{M}\partial_{xx}^2\mathbf{R}\right),\\
\varepsilon^2 \mathbf{J} &= \varepsilon^2 j^n \mathbf{e} - \Delta t A(\partial_x \mathbf{R} + \mathbf{J}),\\
\rho^{n+1} &= \rho^n - \Delta t \tilde{b}^T(\partial_x \mathbf{J} + \mathcal{M}\partial_{xx}^2\mathbf{R}) + \Delta t b^T\left(\mathcal{M}\partial_{xx}^2\mathbf{R}\right),\\
\varepsilon^2 j^{n+1} &= \varepsilon^2 j^n - \Delta t b^T(\partial_x \mathbf{R} + \mathbf{J}),\\
\rho^1 &= \rho_0 \quad j^1 = j_0,\\
j^n(0) &= 0, \quad j^n(1) - \rho^n(1) = -u^n,\\
j^n(0) &= -\rho_x^n(0), \quad j^n(1) = -\rho^n(1)_x.
\end{aligned} \tag{5.7}$$

Using formal computations we derive the (adjoint) equations (5.8) for the Lagrange multipliers $(p^n, q^n)_{n=1}^N$ and the corresponding stage variables $\mathbf{P}, \mathbf{Q}$ with $\mathbf{P} = (P_\ell(\cdot))_{\ell=1}^s$, $P_\ell \in \mathbb{R}^s$ and $\mathbf{Q}$ respectively.

$$\begin{aligned}
p^n &= p^{n+1} + \mathbf{e}^T\mathbf{P}, \qquad\qquad \rho^N - \rho_d - p^N = 0,\\
\varepsilon^2 q^n &= \varepsilon^2 q^{n+1} + \varepsilon^2 \mathbf{e}^T \mathbf{Q}, \qquad \varepsilon^2 q^N = 0,\\
\mathbf{P} &= \Delta t\left(\partial_x(A^T\mathbf{Q}) + \partial_x q^{n+1} b\right) - \Delta t \mathcal{M}\left(\partial_{xx}^2(\tilde{A}^T\mathbf{P}) + \partial_{xx}^2 p^{n+1}\tilde{b}\right)\\
&\quad + \Delta t \mathcal{M}\left(\partial_{xx}^2(A^T\mathbf{P}) + \partial_{xx}^2 p^{n+1} b\right),\\
\varepsilon^2 \mathbf{Q} &= -\Delta t\left(A^T\mathbf{Q} + q^{n+1}b\right) + \Delta t\left(\partial_x(\tilde{A}^T\mathbf{P}) + \partial_x p^{n+1}\tilde{b}\right).
\end{aligned} \tag{5.8}$$

We obtain boundary conditions for (5.8) as

$$q^n(0) = 0, \quad q^n(1) + p^n(1) = 0, \quad q^n(0) = p_x^n(0) \quad \text{and} \quad q^n(1) = p_x^n(1). \tag{5.9}$$

Furthermore, we consider under the assumption of using a type A scheme (we leave on purpose a definitions of these scheme in appendix 5.5) the limit case $\varepsilon = 0$ of the optimal control problem (5.5). Note that for $\varepsilon = 0$ we have $\mathcal{M} = \mathrm{Id}$. The semi–discretized problem is

$$\begin{aligned} &\min \frac{1}{2}\int_0^1 (\rho^N(x)-\rho_d(x))^2 dx + \Delta t \frac{\nu}{2}\sum_{n=1}^{N}(u^n)^2 \\ &\mathbf{R} = \rho^n \mathbf{e} - \Delta t \tilde{A}\left(\partial_x \mathbf{J} + \partial^2_{xx}\mathbf{R}\right) + \Delta t A\left(\partial^2_{xx}\mathbf{R}\right) \\ &\mathbf{J} = -\partial_x \mathbf{R} \\ &\rho^{n+1} = \rho^n - \Delta t \tilde{b}^T\left(\partial_x \mathbf{J} + \partial^2_{xx}\mathbf{R}\right) + \Delta t b^T\left(\partial^2_{xx}\mathbf{R}\right) \\ &\rho^1 = \rho_0 \quad \rho^n_x(0) = 0, \quad \rho^n_x(1) + \rho^n(1) = u^n, \end{aligned} \tag{5.10}$$

and the corresponding adjoint equations are given by (5.11).

$$\begin{aligned} p^n &= p^{n+1} + \mathbf{e}^T\mathbf{P}, \qquad \rho^N - \rho_d - p^N = 0, \\ \mathbf{P} &= \partial_x \overline{\mathbf{Q}} - \Delta t\left(\partial^2_{xx}(\tilde{A}^T\mathbf{P}) + \partial^2_{xx}p^{n+1}\tilde{b}\right) \\ &\quad + \Delta t\left(\partial^2_{xx}(A^T\mathbf{P}) + \partial^2_{xx}p^{n+1}b\right) \\ \overline{\mathbf{Q}} &= \Delta t\left(\partial_x(\tilde{A}^T\mathbf{P}) + \partial_x p^{n+1}\tilde{b}\right) \end{aligned} \tag{5.11}$$

We obtain boundary conditions for (5.11) as

$$p^n_x(1) + p^n(1) = 0, \qquad \text{and} \qquad p^n_x(0) = 0. \tag{5.12}$$

The relation between the limiting problem and the small ε limit of the adjoint equations (5.8) and (5.11) is summarized in the following Lemma.

Lemma 5.2.1 *If the IMEX Runge Kutta method is implicit stiffly accurate (ISA) and of type A, then the $\varepsilon = 0$ limit of (5.8) is given by*

$$\begin{aligned} p^n &= \mathbf{e}^t\mathbf{P}, \;\; \rho^N - \rho_d - p^N = 0, \qquad\qquad q^n = 0, \;\; q^N = 0, \\ \mathbf{P} &= p^{n+1}\mathbf{e}_s + \Delta t \partial_x\left(A^T\mathbf{Q}\right) \\ &\quad - \Delta t\left(\partial^2_{xx}\left(\tilde{A}^T\mathbf{P}\right) - \partial^2_{xx}\left(A^T\mathbf{P}\right)\right) - \Delta t(\tilde{b}^T - \mathbf{e}_s^T\tilde{A})\partial^2_{xx}p^{n+1}\mathbf{e} \\ 0 &= -\Delta t\left(A^T\mathbf{Q} - \partial_x\left(\tilde{A}^T\mathbf{P}\right)\right) + \Delta t(\tilde{b}^T - \mathbf{e}_s\tilde{A})\partial_x p^{n+1}\mathbf{e}^T \end{aligned} \tag{5.13}$$

Further, there exists a linear variable transformation such that a solution to (5.13) is equivalent to a solution of the adjoint equation (5.11) of Problem (5.10) for $\varepsilon = 0$.

Proof 5.2.1 *In the case of implicit stiffly accurateness the IMEX scheme simplifies to*

$$\begin{aligned} \mathbf{R} &= \rho^n\mathbf{e} - \Delta t\tilde{A}(\partial_x\mathbf{J} + \mathcal{M}\partial^2_{xx}\mathbf{R}) + \Delta t A\left(\mathcal{M}\partial^2_{xx}\mathbf{R}\right) \\ \varepsilon^2\mathbf{J} &= \varepsilon^2 j^n\mathbf{e} - \Delta t A(\partial_x\mathbf{R} + \mathbf{J}) \\ \rho^{n+1} &= \mathbf{e}_s^T\mathbf{R} - \Delta t(\tilde{b}^T - \mathbf{e}_s^T\tilde{A})(\partial_x\mathbf{J} + \mathcal{M}\partial^2_{xx}\mathbf{R}), \quad j^{n+1} = \mathbf{e}_s^T\mathbf{J} \end{aligned} \tag{5.14}$$

and the corresponding adjoint equations are given by

$$\begin{aligned}
p^n &= \mathbf{e}^t\mathbf{P},\ \ q^n = \varepsilon^2\mathbf{e}^T\mathbf{Q}, \qquad\qquad \rho^N - \rho_d - p^N = 0,\ \ \varepsilon^2 q^N = 0,\\
\mathbf{P} &= p^{n+1}\mathbf{e}_s + \Delta t\partial_x\left(A^T\mathbf{Q}\right)\\
&\quad - \Delta t\mathcal{M}\left(\partial^2_{xx}\left(\tilde{A}^T\mathbf{P}\right) - \partial^2_{xx}\left(A^T\mathbf{P}\right)\right) - \Delta t(\tilde{b}^T - \mathbf{e}_s^T\tilde{A})\partial^2_{xx}p^{n+1}\mathcal{M}\mathbf{e}\\
\varepsilon^2\mathbf{Q} &= q^{n+1}\mathbf{e}_s - \Delta t\left(A^T\mathbf{Q} - \partial_x\left(\tilde{A}^T\mathbf{P}\right)\right) + \Delta t(\tilde{b}^T - \mathbf{e}_s^T\tilde{A})\partial_x p^{n+1}\mathbf{e}
\end{aligned}$$

Since $\mathcal{M} = \mathrm{Id}$ *in the limit* $\varepsilon = 0$ *we obtain the adjoint equations* (5.13)*. Introducing the transformation*

$$\overline{\mathbf{Q}} = \Delta t A^T\mathbf{Q}.$$

and proceeding yields from (5.13) *the system* (5.15)*.*

$$\begin{aligned}
p^n &= \mathbf{e}^T\mathbf{P},\ \ q^n = 0, \qquad\qquad \rho^N - \rho_d - p^N = 0,\ \ q^N = 0,\\
\mathbf{P} &= p^{n+1}\mathbf{e}_s + \partial_x\overline{\mathbf{Q}}\\
&\quad - \Delta t\left(\partial^2_{xx}\left(\tilde{A}^T\mathbf{P}\right) - \partial^2_{xx}\left(A^T\mathbf{P}\right)\right) - \Delta t(\tilde{b}^T - \mathbf{e}_s^T\tilde{A})\partial^2_{xx}p^{n+1}\mathbf{e}\\
\overline{\mathbf{Q}} &= \Delta t\partial_x\left(\tilde{A}^T\mathbf{P}\right) + \Delta t(\tilde{b}^T - \mathbf{e}_s^T\tilde{A})\partial_x p^{n+1}\mathbf{e}
\end{aligned}\tag{5.15}$$

The latter are the adjoint equations (5.11) *to problem* (5.10) *provided an implicit stiffly accurate scheme* (5.14) *has been used. Therein* $\rho^{n+1} = \rho^n - \Delta t\tilde{b}^T\left(\partial_x\mathbf{J} + \partial^2_{xx}\mathbf{R}\right) + \Delta t b^T\left(\partial^2_{xx}\mathbf{R}\right)$ *becomes* $\rho^{n+1} = \mathbf{e}_s\mathbf{R} - \Delta t\left(\tilde{b}^T - \mathbf{e}_s\tilde{A}\right)\left(\partial_x\mathbf{J} + \partial^2_{xx}\mathbf{R}\right)$*, which yields further simplifications in* (5.13) *and* (5.15)*, respectively.*

A particular, yet important case of Lemma 5.2.1 are the so–called globally stiffly accurate IMEX scheme. They fulfil additionally $(\tilde{b}^T - \mathbf{e}_s^T\tilde{A}) = 0$.

5.3 Optimal choice of $\mathcal{M}$

In the following section we discuss the optimal choice of μ in equation (5.6). We want to avoid parabolic stiffness for small value of ε, and the numerical instabilities due to the discretization of the term $(j + \mu\rho_x)_x$. In [38] the following formula has been used $\mu = \exp(-\varepsilon/\Delta x)$, here we want to choose μ in such a way that Chapman-Enskog expansion with respect to ε at least to order $O(\varepsilon^2)$ and the term $j + \mu\rho_x$ vanishes. It can been shown that independent of μ a stiffly accurate asymptotic-preserving IMEX yields an asymptotic–preserving scheme for the limit equation.

Considering an $s-$stage IMEX scheme and a semi–discretization of (5.5) as in (5.7), the optimal choice of an diagonal matrix $\mathcal{M}$, such that the explicit term $\mathbf{J} + \mathcal{M}\partial_x\mathbf{R}$ vanishes in the $O(\varepsilon^2)$ regime is presented in the following lemma.

Lemma 5.3.1 *If the IMEX scheme is of type A an optimal choice for $\mathcal{M}$ in the $O(\varepsilon^2)$ regime for scheme*

$$\begin{aligned}
\mathbf{R} &= \rho^n \mathbf{e} - \Delta t \tilde{A} \partial_x \left(\mathbf{J} + \mathcal{M}\partial_x \mathbf{R}\right) + \Delta t A \mathcal{M} \partial^2_{xx} \mathbf{R} \\
\varepsilon^2 \mathbf{J} &= \varepsilon^2 j^n \mathbf{e} - \Delta t A (\partial_x \mathbf{R} + \mathbf{J}) \\
\rho^{n+1} &= \rho^n - \Delta t \tilde{b}^T \partial_x \left(\mathbf{J} + \mathcal{M}\partial_x \mathbf{R}\right) + \Delta t b^T \mathcal{M} \partial^2_{xx} \mathbf{R} \\
\varepsilon^2 j^{n+1} &= \varepsilon^2 j^n - \Delta t b^T (\partial_x \mathbf{R} + \mathbf{J})
\end{aligned} \tag{5.16}$$

is given by

$$\mathcal{M} = \Delta t \left(\varepsilon^2 Id + \Delta t \ \operatorname{diag}(A)\right)^{-1} \ \operatorname{diag}(A). \tag{5.17}$$

The formula follows straightforward substituting stage by stage the approximation of order $O(\varepsilon^2)$ in the subsequent stages

$$\begin{aligned}
J_1 &= -\frac{a_{11}\Delta t}{\varepsilon^2 + a_{11}\Delta t}\partial_x R_1 + O(\varepsilon^2), \\
J_2 &= -\frac{a_{22}\Delta t}{\varepsilon^2 + a_{22}\Delta t}\partial_x R_2 - \frac{a_{21}\Delta t}{\varepsilon^2 + a_{22}\Delta t}\underbrace{(\partial_x R_1 + J_1)}_{O(\varepsilon^2)} + O(\varepsilon^2) = -\frac{a_{22}\Delta t}{\varepsilon^2 + a_{22}\Delta t}\partial_x R_2 + O(\varepsilon^2) \\
&\vdots \\
J_i &= -\frac{a_{ii}\Delta t}{\varepsilon^2 + a_{ii}\Delta t}\partial_x R_i - \sum_{j=1}^{i-1}\frac{a_{ij}\Delta t}{\varepsilon^2 + a_{ii}\Delta t}\underbrace{(\partial_x R_j + J_j)}_{O(\varepsilon^2)} + O(\varepsilon^2) = -\frac{a_{ii}\Delta t}{\varepsilon^2 + a_{ii}\Delta t}\partial_x R_i + O(\varepsilon^2).
\end{aligned}$$

We leave a rigorous proof in subsection 5.6.

Remark 5.3.1 *Note that $\mathcal{M} = \operatorname{diag}(\mu_j^n)$ is not depending on t_n, i.e, $\mu_j^n \equiv \mu_j$, and the solution of (5.17) can be computed for the stages once and for all. Moreover (5.17) tells us that when $\varepsilon \to 0$, $\mathcal{M}$ has the expected behavior, namely $\mathcal{M} \to \mathrm{Id}$.*

5.4 Numerical results

For the temporal discretization we use different IMEX schemes fulfilling the properties of Lemma 5.2.1. We consider second–order in time schemes. The IMEX GSA(3,4,2), [107], as given by the Butcher tables in table 5.4 is a globally stiffly accurate scheme which is of type A. The implicit part is invertible and the last row of implicit and explicit scheme coincide. It is of second–order as the numerical results show. Further, we consider the second–order IMEX SSP(3,3,2) scheme, [38], (table 5.5) which is only implicitly stiffly accurate and of type A. In view of Theorem 3.1 [107] we observe that SSP(3,3,2) is symplectic. Theorem 2.1 [107] guarantees that for all considered schemes the convergence order of the IMEX scheme applied to the optimality system is also of second–order.

For the spatial discretization we introduce an equidistant grid with M grid points $\{x_i\}_{i=1}^M$ and grid size Δx, such that $x_1 = \frac{\Delta x}{2}$ and $x_M = 1 - \frac{\Delta x}{2}$. We set $\rho^n(x_i) = \rho_i^n$ and $j^n(x_i) = j_i^n$.

Since the Goldstein-Taylor model depends on ε, we expect parabolic behavior for $\varepsilon \ll 1$ and hyperbolic behavior else. We use second order central difference for the diffusive part ρ_{xx} and hyperbolic discretization based on an Upwind scheme for the advective terms. In order to determine the Upwind direction, we recall from section 5.1 the definition of the macroscopic variables

$$\rho = f^+ + f^-, \qquad j = \frac{1}{\varepsilon}(f^+ - f^-). \tag{5.18}$$

We obtain for f^+, the density of particles with positive velocity, the Upwind scheme,

$$\frac{f_i^+ - f_{i-1}^+}{\Delta x} = \frac{f_{i+1}^+ - f_{i-1}^+}{2\Delta x} - \frac{\Delta x}{2}\frac{f_{i+1}^+ - 2f_i^+ - f_{i-1}^+}{(\Delta x)^2}.$$

Similar for the scheme of f^-. By combining the discretization for f^+ and f^- we obtain the discrete stencils in the original variables by applying (5.18), as follows:

$$D^h\rho = D^c\rho - \frac{\varepsilon\Delta x}{2}D^2 j, \qquad D^h j = D^c j - \frac{\Delta x}{2\varepsilon}D^2\rho \tag{5.19}$$

where D^c is the stencil for central difference $\frac{1}{\Delta x}\begin{pmatrix}-1 & 0 & 1\end{pmatrix}$ and D^2 the second order central difference $\frac{1}{(\Delta x)^2}\begin{pmatrix}1 & -2 & 1\end{pmatrix}$. Using a convex combination of the discretization of the diffusive term with the hyperbolic part by the function $\Phi = \Phi(\varepsilon)$ we finally obtain

$$D\rho = \Phi D^c\rho + (1-\Phi)D^h\rho, \qquad Dj = \Phi D^c j + (1-\Phi)D^h j \tag{5.20}$$

The function Φ is chosen such that $\Phi(0) = 1$ and $\Phi(\varepsilon)\frac{\Delta x}{2\varepsilon} \to 0$ for $\varepsilon \to 1$. The simplest possible way is $\Phi = 1 - \varepsilon$, but more cleaver choices have been proposed in [**?**], where the value of Φ coincides with $\mu = \exp(-\varepsilon/\Delta x)$ or in [**?**] with $\Phi = 1 - \tanh(\varepsilon/\Delta x)$.

In all cases we discretize the with a spatial grid size $\Delta x \approx \Delta t$ since we avoid the parabolic CFL condition due to introduced splitting, (5.6). The discretization of $\Delta x \approx \Delta t$ is the typical hyperbolic CFL type condition induced by the transport.

5.4.1 Order analysis

To verify the theoretical results numerically we set up the following test problem. We consider the parabolic case. Let $\varepsilon = 0, \nu = 0, u_l = -1$ and $u_r = 1$. Further

N	$\lVert\rho_N^* - \rho_d\rVert_{L^\infty(L^1(\Omega))}$	$\lVert\rho_N^* - \rho_d\rVert_{L^\infty(L^\infty(\Omega))}$	$\lVert p_N^*\rVert_{L^\infty(L^1(\Omega))}$	$\lVert p_N^*\rVert_{L^\infty(L^\infty(\Omega))}$
20	1.31e-04	2.69e-07	1.25e-04	2.02e-07
40	3.10e-05 (2.08)	6.08e-08 (2.14)	3.03e-05 (2.04)	4.88e-08 (2.04)
80	7.55e-06 (2.04)	1.43e-08 (2.09)	7.46e-06 (2.02)	1.21e-08 (2.01)
160	1.86e-06 (2.01)	3.45e-09 (2.05)	1.85e-06 (2.01)	3.01e-09 (2.00)
320	4.62e-07 (2.00)	8.41e-10 (2.03)	4.61e-07 (2.00)	7.55e-10 (1.99)

Table 5.1: Order results for the GSA(3,4,2), table 5.4, $\varepsilon = 0$. In brackets the the $\log_2$-ratio between the results from two subsequent step width.

set $\rho_0 = cos(x)$, $j_0 = 0$ and $\rho_d(x) = e^{-T}\cos(x)$. Then, the solution to the optimal control problem (5.4) – (5.5) is given analytically by $u^*(t) = e^{-t}(\cos(1) - \sin(1))$ and $J = 0$. Within this setting $\rho(x,t) = e^{-t}\cos(x)$ is solution of (5.5) and $p^*(t,1) = 0$. The domain is $\Omega = [0,1]$ and the terminal time $T = 1$.

We compute the numerical solution for different values of $N \in \{20, 40, 80, 160, 320\}$ using different IMEX schemes. We denote by ρ_N^* and p_N^* the solution to (5.8) with initial values $\rho_d = \rho(x,T)$ and $\rho^N = \rho_N^*$. We compare ratios of L^∞ and L^1 errors of the approximate solutions using the following norms:

$$L^\infty(L^1(\Omega)) := L^\infty(0,T;L^1(\Omega)) \qquad \textit{and} \qquad L^\infty(L^\infty(\Omega)) := L^\infty(0,T;L^\infty(\Omega)).$$

The results for different IMEX schemes are listed in table 5.1 and 5.2.

N	$\lVert\rho_N^* - \rho_d\rVert_{L^\infty(L^1(\Omega))}$	$\lVert\rho_N^* - \rho_d\rVert_{L^\infty(L^\infty(\Omega))}$	$\lVert p_N^*\rVert_{L^\infty(L^1(\Omega))}$	$\lVert p_N^*\rVert_{L^\infty(L^\infty(\Omega))}$
20	1.29e-04	2.73e-07	1.22e-04	1.96e-07
40	3.07e-05 (2.06)	6.15e-08 (2.15)	2.99e-05 (2.03)	4.80e-08 (2.02)
80	7.51e-06 (2.03)	1.44e-08 (2.09)	7.42e-06 (2.02)	1.19e-08 (2.00)
160	1.85e-06 (2.01)	3.46e-09 (2.05)	1.85e-06 (2.00)	3.00e-09 (1.99)
320	4.62e-07 (2.00)	8.43e-10 (2.03)	4.61e-07 (2.00)	7.52e-10 (1.99)

Table 5.2: Order results for the SSP2(3,3,2), table 5.5, $\varepsilon = 0$. In brackets the $\log_2$-ratio between the results from two subsequent step width corresponds.

As expected we observe the convergence order of two for all discussed schemes. We tested the example for stiffly accurate (SSP2(3,3,2)) as well as globally stiffly accurate schemes (GSA(3,4,2)). The order two is in particular preserved in the limit $\varepsilon = 0$ as expected by the previous Lemmas.

5.4.2 Computational results on the optimal control problem

We compare the IMEX methods applied to the Goldstein–Taylor model in the limit case $\varepsilon = 0$ with the numerical solution presented in [156]. Therein, the limit problem has been studied using parameters: $T = 1.58, \rho_o = j_0 = 0, \rho_d(x) = 0.5(1 - x^2), \nu = 0.001, u_l = -1$ and $u_r = 1$. We furthermore set $N = 100$ and $M = 50$. We use a gradient based optimization to iteratively compute the optimal control u^* using an implicit stiffly accurate scheme (ISA). The numerical approximation to the gradient for the reduced objective functional $\tilde{J}(u)$ is then given by

$$\nabla \tilde{J} = \Delta t(\nu u^n + p^n)$$

where p^n is the solution to the adjoint equation (5.8), respectively (5.13), at time t^n. The terminal condition for the gradient based optimization is $\|proj_{[u_l,u_r]}(\nabla \tilde{J})\|_{L^2(0,T)} \leqslant 10^{-6}$.

IMEX	$\varepsilon = 0$	$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 0.8$	$\varepsilon = 1$
GSA(3,4,2)	$6.51 \cdot 10^{-4}$	$5.94 \cdot 10^{-4}$	$2.85 \cdot 10^{-4}$	$2.47 \cdot 10^{-4}$	$2.44 \cdot 10^{-4}$
SSP(3,3,2)	$6.52 \cdot 10^{-4}$	-	$2.84 \cdot 10^{-4}$	$2.46 \cdot 10^{-4}$	$2.43 \cdot 10^{-4}$

Table 5.3: Results for $J(u_\varepsilon^*)$, different IMEX schemes and values of ε. For $\varepsilon = 0$ in [156], they obtain $J(u) = 6.86 \cdot 10^{-4}$.

The final values for $J(u_\varepsilon^*)$ for the different schemes are presented in table 5.3. The calculated values with our method $J(u_0^*)$ are consistent with respect to the numerical discretization in space and time to the ones in [156]. Note that in the limit $\varepsilon = 0$ we do not have a parabolic CFL condition due to the applied splitting and the obtained results are precisely as in [156].

Figure 5.1 shows the numerical solutions using GSA(3,4,2) scheme for different values of $\varepsilon \in \{0, 0.1, 0.5, 1\}$. The globally stiffly accurate IMEX schemes yield solutions to the $\varepsilon-$dependent class of optimization problems (5.5) across the full range of parameters ε.

In figure 5.2 we plot the numerical solutions using SSP(3,3,2) scheme for different values of $\varepsilon \in \{0, 0.5, 0.8, 1\}$. As in [38] shown, SSP(3,3,2) is not globally stiffly accurate, and therefore we cannot expect stability for small values of ε, even if $\varepsilon = 0$ provides a stable solution. Further, we set $N = 200$ for similar reasons.

In both figures, one can observe oscillations at the boundary $x = 1$ for values of $\varepsilon > 0.25$. This is caused due to the assumption $j = -\rho_x$ in $x = 1$, which holds true just for $\varepsilon = 0$. We set $\Phi = 0.3$ for GSA(3,4,2) and $\varepsilon = 1$. Further for SSP(3,3,2) and $\varepsilon = 0.5$ we set $\Phi = 0.385$. All other values of ε are treated with $\Phi = 1 - \varepsilon^3$.

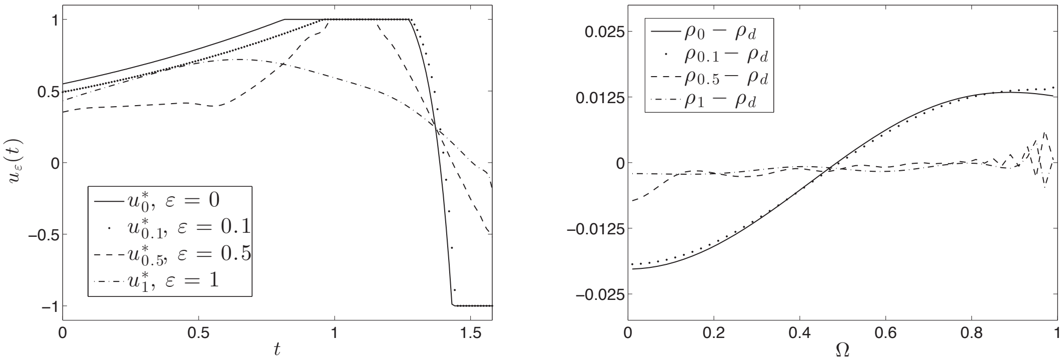


Figure 5.1: Numerical solution, using GSA(3,4,2) scheme, see appendix 5.5, for 150 time steps and 50 grid points in space. The left part of the plot shows the optimal controls u_ε^* for different values of ε. On the right plot we show the corresponding optimal states $\rho_\varepsilon^*(\cdot, T) - \rho_d$ for different choices of ε.

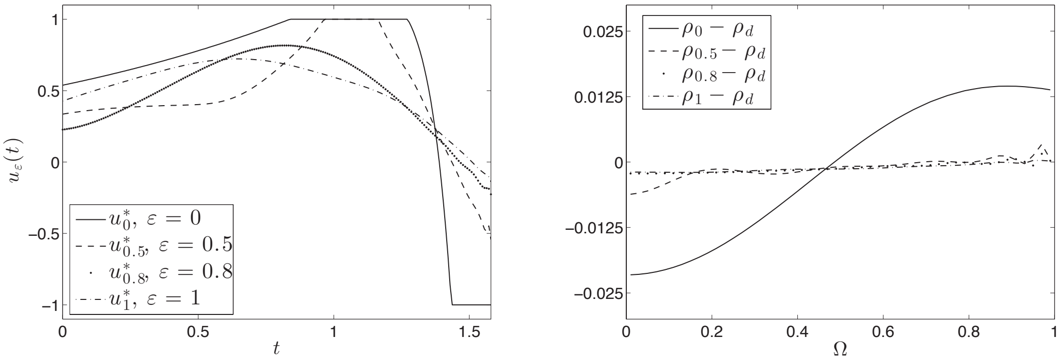


Figure 5.2: Numerical solution, using SSP2(3,3,2), i.e. table 5.5, for 200 time steps and 50 grid points in space. On the left the optimal control u^* is plotted. The right part shows the difference of the optimal state to the desired state, i.e. $\rho_\varepsilon^*(\cdot, T) - \rho_d$.

5.5 Definitions of implicit–explicit Runge–Kutta methods

We consider the Cauchy problem for a system of ODEs such that

$$y' = f(y) + g(y), \qquad y(0) = y_0, \qquad t \in [0, T], \tag{5.21}$$

where $y(t) \in \mathbb{R}$ and $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ Lipschitz continuous functions. Using an Implicit-Explict Runge–Kutta method with time step Δt we obtain the following numerical scheme for (5.21)

$$\begin{aligned}\mathbf{Y} &= y^n\mathbf{e} + \Delta t\left(\tilde{A}\mathbf{F}(\mathbf{Y}) + A\mathbf{G}(\mathbf{Y})\right)\\ y^{n+1} &= y^n + \Delta t\left(\tilde{b}^T\mathbf{F}(\mathbf{Y}) + b^T\mathbf{G}(\mathbf{Y})\right),\end{aligned}$$

where $\mathbf{Y} = (Y_l(\cdot))_{l=1}^s$ denotes the s stage variables, and $\mathbf{F}(\mathbf{Y}^n) = (f(Y_l))_{l=1}^s$, $\mathbf{G}(\mathbf{Y}) = (g(Y_l))_{l=1}^s$, moreover $\mathbf{e} = (1,\ldots,1) \in \mathbb{R}^s$. The matrices $\tilde{A}, A$ are $s \times s$ matrices, and $\tilde{c}, c, \tilde{b}, b \in \mathbb{R}^s$. We take in account IMEX schemes satisfying the following definition

Definition 5.5.1 *A diagonally implicit IMEX Runge Kutta (DIRK) method is such that matrices $\tilde{A}$, and A are lower triangular, where $\tilde{A}$ has zero diagonal.*

Further we consider the following basic assumptions on $\tilde{c}, c, \tilde{b}, b \in \mathbb{R}^s$

$$\sum_{i=1}^{s} b_i = 1, \qquad \sum_{i=1}^{s} \tilde{b}_i = 1, \qquad \tilde{c}_i = \sum_{j=1}^{i-1} \tilde{a}_{ij}, \qquad c_i = \sum_{j=1}^{i} a_{ij},$$

Those conditions need to be fulfilled for a first order Runge Kutta method. Increasing the order of a Runge Kutta method increases the number of restrictions on the coefficients in the Butcher tables. For IMEX methods up to order $k = 3$, the number of constraints can be reduced if $c = \tilde{c}$ and $b = \tilde{b}$, [152, 151].

Definition 5.5.2 (Type A [38, 145]) *If A is invertible the IMEX scheme is of* **type A**.

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0\\ 3/2 & 3/2 & 0 & 0 & 0\\ 1/2 & 5/6 & -1/3 & 0 & 0\\ 1 & 1/3 & 1/6 & 1/2 & 0\\ \hline & 1/3 & 1/6 & 1/2 & 0 \end{array} \qquad \begin{array}{c|cccc} 1/2 & 1/2 & 0 & 0 & 0\\ 5/4 & 3/4 & 1/2 & 0 & 0\\ 1/4 & -1/4 & 0 & 1/2 & 0\\ 1 & 1/6 & -1/6 & 1/2 & 1/2\\ \hline & 1/6 & -1/6 & 1/2 & 1/2 \end{array}$$

Table 5.4: GSA(3,4,2), [107], Type A scheme and globally stiffly accurate, (GSA).

Definition 5.5.3 (Type GSA [38]) *An IMEX method is* **globally stiffly accurate** *(***GSA***), if $\tilde{c}_s = c_s = 1$ and*

$$\tilde{b}^T = \mathbf{e}_s^T\tilde{A} \quad \textit{and} \quad b^T = \mathbf{e}_s^T A, \tag{5.22}$$

If the previous equalities hold only for the implicit part, the method is **implicit stiffly accurate** *(***ISA***).*

To denote each IMEX scheme we use the following convention for the names of the schemes: Acronym(σ_E, σ_I, k),where σ_E denoting the effective number of stages of the explicit, σ_I of the implicit scheme. and k the combined order of accuracy.

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 \\
1 & 1/2 & 1/2 & 0 \\
\hline
 & 1/3 & 1/3 & 1/3
\end{array}
\qquad
\begin{array}{c|ccc}
1/4 & 1/4 & 0 & 0 \\
1/4 & 0 & 1/4 & 0 \\
1 & 1/3 & 1/3 & 1/3 \\
\hline
 & 1/3 & 1/3 & 1/3
\end{array}
$$

Table 5.5: SSP2(3,3,2) [145], Type A and implicit stiffly accurate scheme (ISA).

5.6 Proof of Lemma 5.3.1

Let consider system (5.16), we can decompose matrix A in this way $A = D + L$, where $D = \text{diag}(A)$ and L is the lower triangular part of A, Therefore we can rewrite the second equation for $\mathbf{J}$ in this way

$$
\begin{aligned}
\varepsilon^2\mathbf{J} =& \varepsilon^2 j^n\mathbf{e} - \Delta t D\left(\partial_x\mathbf{R} + \mathbf{J}\right) - \Delta t L\left(\partial_x\mathbf{R} + \mathbf{J}\right) \\
\left(\varepsilon^2 Id + \Delta t D\right)\mathbf{J} =& \varepsilon^2 j^n\mathbf{e} - \Delta t D\partial_x\mathbf{R} - \Delta t L\left(\partial_x\mathbf{R} + \mathbf{J}\right)
\end{aligned}
$$

Neglecting the $O(\varepsilon^2)$ term and inverting the diagonal matrix on the lefthand side we have

$$
\mathbf{J} = -\underbrace{\Delta t\left(\varepsilon^2 Id + \Delta t D\right)^{-1} D}_{\mathcal{M}}\,\partial_x\mathbf{R} - \underbrace{\Delta t\left(\varepsilon^2 Id + \Delta t D\right)^{-1} L}_{\mathcal{K}}\left(\partial_x\mathbf{R} + \mathbf{J}\right) + o(\varepsilon^2). \tag{5.23}
$$

Recursively we substitute in J (5.23) itself, the first recursion gives

$$
\mathbf{J} = -\mathcal{M}\partial_x\mathbf{R} + \mathcal{K}\left(Id - \mathcal{M}\right)\partial_x\mathbf{R} + \mathcal{K}^2\left(\partial_x\mathbf{R} + \mathbf{J}\right) + o(\varepsilon^2)
$$

applying the recursion $s-1$ times we obtain

$$
\begin{aligned}
\mathbf{J} =& -\mathcal{M}\partial_x\mathbf{R} - \sum_{l=1}^{s-1}(-\mathcal{K})^l\left(Id - \mathcal{M}\right)\partial_x\mathbf{R} - (-\mathcal{K})^s\left(\partial_x\mathbf{R} + \mathbf{J}\right) + o(\varepsilon^2) = \\
=& -\mathcal{M}\partial_x\mathbf{R} + \left(\sum_{l=1}^{s-1}(-1)^{l-1}\mathcal{K}^l\right)\left(Id - \mathcal{M}\right)\partial_x\mathbf{R} + o(\varepsilon^2),
\end{aligned}
$$

where in the last equation $\mathcal{K}^s$ vanishes since it is a nilpotent matrix of grade s, moreover each element of matrix $Id - \mathcal{M}$ has order $o(\varepsilon^2)$, from a direct computation on the general i element of the diagonal matrix we have

$$
(Id - \mathcal{M})_i = 1 - \frac{\Delta t a_{ii}}{\varepsilon^2 + \Delta t a_{ii}} = 1 - \frac{\Delta t a_{ii}}{\varepsilon^2 + \Delta t a_{ii}} = \frac{\varepsilon^2}{\varepsilon^2 + \Delta t a_{ii}}
$$

Thus the expression for $\mathbf{J}$ reads

$$\mathbf{J} = -\,\mathcal{M}\partial_x \mathbf{R} + o(\varepsilon^2),$$

which cancel exactly the explicit part of the semi–discretize scheme, in the $o(\varepsilon^2)$ regime. Hence the appropriate choice for $\mathcal{M}$ is given by

$$\mathcal{M} = \Delta t \left(\varepsilon^2 Id + \Delta t D\right)^{-1} D. \tag{5.24}$$

In table 5.6 we show $\mathcal{M}$ for different schemes using the provided method. Note that we use the same $\mathcal{M}$ for the adjoint equations.

IMEX	$\mathcal{M}(\varepsilon)$
GSA(3,4,2)	$\frac{\Delta t}{2\,\varepsilon^2+\Delta t}\,\mathrm{diag}(1,1,1,1)$
SSP2(3,3,2)	$\mathrm{diag}\left(\frac{\Delta t}{4\,\varepsilon^2+\Delta t}, \frac{\Delta t}{4\,\varepsilon^2+\Delta t}, \frac{\Delta t}{3\varepsilon^2+\Delta t}\right)$

Table 5.6: Optimal choice of matrix $\mathcal{M}$ for the different IMEX schemes used.

5.7 Conclusions

In this chapter we develop an asymptotic preserving implicit-explicit Runge-Kutta scheme for optimal control problems of boundary problems governed by the *Goldstein–Taylor* model. We investigated the relation of time integration schemes and the formal Chapman-Enskog type limiting procedure. For the class of stiffly accurate implicit-explicit Runge-Kutta methods (IMEX) the discrete optimality system also provides a stable numerical method for optimal control problems governed by the heat equation. The stability of the method is optimized for the stiff regime thanks to the BPR approach and an optimal choice of the related function $\mu(\varepsilon)$, which permit a fully implicit solver for the limiting heat equation and numerically cancel any loss of accuracy due to the explicit part of the method. The methodology presented opens new perspectives to extend the same techniques for radiative transfer equation, which gives a description of radiotherapy process in biomedical application, results are a under investigations.

Bibliography

[1] M. Agueh, R. Illner, and A. Richardson. Analysis and simulations of a refined flocking and swarming model of cucker-smale type. *Kinetic and Related Models*, 4(1):1–16, 2011.

[2] G. Albi, D. Balagué, J. A. Carrillo, and J. von Brecht. Stability analysis of flock and mill rings for 2nd order models in swarming. *submitted in revised form to SIAM J. App. Math.*, 2013.

[3] G. Albi, M. Herty, C. Jörres, and L. Pareschi. Asymptotic preserving time-discretization of optimal control problems for the Goldstein-Taylor model. *submitted*, 2013.

[4] G. Albi, M. Herty, C. Jörres, and L. Pareschi. Asymptotic preserving schems for the optimal control of radiative transfer, application to radiotherapy. *in preparation*, 2014.

[5] G. Albi, M. Herty, and L. Pareschi. Kinetic description of optimal control problems in consensus modeling. *submitted to Comm. Math. Sci.*, 2013.

[6] G. Albi and L. Pareschi. Binary interaction algorithms for the simulation of flocking and swarming dynamics. *Multiscale Model. Simul.*, 11(1):1–29, 2013.

[7] G. Albi and L. Pareschi. Modeling of self-organized systems interacting with a few individuals: from microscopic to macroscopic dynamics. *Appl. Math. Lett.*, 26(4):397–401, 2013.

[8] G. Aletti, G. Naldi, and G. Toscani. First-order continuous models of opinion formation. *SIAM J. Appl. Math.*, 67(3):837–853 (electronic), 2007.

[9] X. Antoine and M. Lemou. Wavelet approximations of a collision operator in kinetic theory. *C. R. Math. Acad. Sci. Paris*, 337(5):353–358, 2003.

[10] I. Aoki. A simulation study on the schooling mechanism in fish. *Bull. Japan Soc. Sci. Fish*, 48:1081–1088, 1982.

[11] D. Armbruster and C. Ringhofer. Thermalized kinetic and fluid models for re-entrant supply chains. *SIAM J. Multiscale Modeling and Simulation*, 3(782–800), 2005.

[12] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25(2-3):151–167, Nov. 1997.

[13] W. R. Ashby. Principles of the self-organizing dynamic system. *Journal of General Psychology*, 37:125–128, 1947.

[14] W. R. Ashby. Principles of the self-organizing system. *Principles of Self-Organization: Transactions of the University of Illinois Symposium*, pages 255–278, 1962.

[15] M. Aureli, F. Fiorilli, and M. Porfiri. Portraits of self-organization in fish schools interacting with robots. *Physica D: Nonlinear Phenomena*, 241(9):908 – 920, 2012.

[16] H. Babovsky. On a simulation scheme for the Boltzmann equation. *Math. Methods Appl. Sci.*, 8(2):223–233, 1986.

[17] D. Balagué, J. A. Carrillo, T. Laurent, and G. Raoul. Dimensionality of local minimizers of the interaction energy. *to appear in ARMA*, 2013.

[18] D. Balagué, J. A. Carrillo, T. Laurent, and G. Raoul. Nonlocal interactions by repulsive-attractive potentials: radial ins/stability. *to appear in Physica D*, 2013.

[19] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1232–1237, 2008.

[20] M. K. Banda and M. Herty. Adjoint IMEX-based schemes for control problems governed by hyperbolic conservation laws. *Comput. Optim. Appl.*, 51(2):909–930, 2012.

[21] A. B. T. Barbaro, K. Taylor, P. F. Trethewey, L. Youseff, and B. Birnir. Discrete and continuous models of the dynamics of pelagic fish: application to the capelin. *Math. Comput. Simulation*, 79(12):3397–3414, 2009.

[22] N. Bellomo, G. Ajmone Marsan, and A. Tosin. *Complex Systems and Society. Modeling and Simulation.* SpringerBriefs in Mathematics. Springer, 2013.

[23] N. Bellomo, M. A. Herrero, and A. Tosin. On the dynamics of social conflicts: looking for the black swan. *Kinet. Relat. Models*, 6(3):459–479, 2013.

[24] N. Bellomo and J. Soler. On the mathematical theory of the dynamics of swarms viewed as complex systems. *Math. Models Methods Appl. Sci.*, 22(suppl. 1):1140006, 29, 2012.

[25] E. Ben-Naim. Opinion dynamics, rise and fall of political parties. *Europhys. Lett.*, 69:671, 2005.

[26] G. Beni. From swarm intelligence to swarm robotics. In E. Sahin and W. Spears, editors, *Swarm Robotics*, volume 3342 of *Lecture Notes in Computer Science*, pages 1–9. Springer Berlin Heidelberg, 2005.

[27] A. Bensoussan, J. Frehse, and P. Yam. *Mean Field Games and Mean Field Type Control Theory.* Series: SpringerBriefs in Mathematics, New York, 2013.

[28] A. Bertozzi, J. von Brecht, H. Sun, T. Kolokolnikov, and D. Uminsky. Ring patterns and their bifurcations in a nonlocal model of biological swarms. *Preprint.*

[29] A. L. Bertozzi, H. Sun, J. von Brecht, T. Kolokolnikov, and D. Uminsky. Ring patterns and their bifurcations in the model of biological swarms. *Submitted.*

[30] G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms. I. *Comm. Pure Appl. Math.*, 44(2):141–183, 1991.

[31] S. Bianchini. On the shift differentiability of the flow generated by a hyperbolic system of conservation laws. *Discrete Contin. Dynam. Systems*, 6:329–350, 2000.

[32] G. Bird. Approach to translational equilibrium in a rigid sphere gas. *Physics of Fluids*, 6:1518, 1963.

[33] B. Birnir. An ODE model of the motion of pelagic fish. *J. Stat. Phys.*, 128(1 - 2):535–568, 2007.

[34] A. Bobylev and K. Nanbu. Theory of collision algorithms for gases and plasmas based on the boltzmann equation and the Landau-Fokker-Planck equation. *Physical Review E*, 61(4):4576, 2000.

[35] E. Bonabeau, M. Dorigo, and G. Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems.* Oxford University Press, New York, 1999.

[36] J. F. Bonnans and J. Laurent-Varin. Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control. *Numer. Math.*, 103(1):1–10, 2006.

[37] A. Borzì and S. Wongkaew. Modeling and control through leadership of a refined flocking system. *preprint*, 2014.

[38] S. Boscarino, L. Pareschi, and G. Russo. Implicit-explicit runge–kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit. *SIAM J. Scient. Comp.*, 35(1):A22–A51, 2013.

[39] L. Boudin and F. Salvarani. A kinetic approach to the study of opinion formation. *ESAIM: Math. Mod. Num. Anal.*, 43:507–522, 2009.

[40] L. Breger, J. How, and A. Richards. Model predictive control of spacecraft formations with sensing noise. In *American Control Conference, 2005. Proceedings of the 2005*, pages 2385–2390 vol. 4, June 2005.

[41] A. Bressan and G. Guerra. Shift-differentiability of the flow generated by a conservation law. 1995.

[42] A. Bressan and M. Lewicka. Shift differentials of maps in BV spaces. Number 401. Chapman & Hall/CRC, Boca Raton, FL, Boca Raton, 1999.

[43] R. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta numerica*, 1998:1–49, 1998.

[44] E. Camacho and C. Bordons. *Model predictive control*. Springer, USA, 2004.

[45] S. Camazine, J. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. *Self-Organization in Biological Systems*. Princeton University Press, Princeton, 2001.

[46] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[47] J. A. Cañizo, J. A. Carrillo, and J. Rosado. A well-posedness theory in measures for some kinetic models of collective motion. *Math. Models Methods Appl. Sci.*, 21(3):515–539, 2011.

[48] M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and control of alignment models. *submitted*, 2012.

[49] M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and optimal control of the Cucker-Smale model. *Math. Control Relat. Fields*, 3(4):447–466, 2013.

[50] J. Carrillo, M. D'Orsogna, and V. Panferov. Double milling in self-propelled swarms from kinetic theory. *Kin. Rel. Mod.*, 2:363–378, 2009.

[51] J. Carrillo, A. Klar, S. Martin, and S. Tiwari. Self-propelled interacting particle systems with roosting force. *Math. Models Methods Appl. Sci*, 20:1533–1552, 2010.

[52] J. Carrillo, S. Martin, and V. Panferov. A new interaction potential for swarming models. *Physica D: Nonlinear Phenomena*, 260(0):112 – 126, 2013. Emergent Behaviour in Multi-particle Systems with Non-local Interactions.

[53] J. A. Carrillo, M. Fornasier, J. Rosado, and G. Toscani. Asymptotic flocking dynamics for the kinetic Cucker-Smale model. *SIAM J. Math. Anal.*, 42:218–236, 2010.

[54] J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. Particle, kinetic, and hydrodynamic models of swarming. pages 297–336, 2010.

[55] J. A. Carrillo, Y. Huang, and S. Martin. Nonlinear stability of flock solutions in second-order swarming models. *preprint*, 2013.

[56] J. A. Carrillo, Y. Huang, and S. Martin. Stability of flock solutions in second-order swarming models. *work in preparation*, 2013.

[57] C. Castro, F. Palacios, and E. Zuazua. An alternating descent method for the optimal control of the inviscid Burgers equation in the presence of shocks. *Math. Models Methods Appl. Sci.*, 18:369–416, 2008.

[58] Y. Chuang, M. R. D'Orsogna, D. Marthaler, and L. C. A. Bertozzi. State transitions and the continuum limit for interacting, self-propelled particles. *Phys. D*, 232:33–47, 2007.

[59] R. M. Colombo and M. Lecureux-Mercier. An analytical framework to describe the interactions between individuals and a continuum. *Journal of Nonlinear Science*, 22(1):39–61, 2012.

[60] R. M. Colombo and N. Pogodaev. Confinement strategies in a model for the interaction between individuals and a continuum. *SIAM J. Appl. Dyn. Syst.*, 11(2):741–770, 2012.

[61] R. M. Colombo and N. Pogodaev. On the control of moving sets: positive and negative confinement results. *SIAM J. Control Optim.*, 51(1):380–401, 2013.

[62] S. Cordier, L. Pareschi, and G. Toscani. On a kinetic model for a simple market economy. *J. Stat. Phys.*, 120(1-2):253–277, 2005.

[63] P. A. Corning. The re-emergence of emergence: A venerable concept in search of a theory. *Complexity*, 7(6):18–30, 2002.

[64] I. Couzin, J. Krause, N. Franks, and S. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433:513–516, 2005.

[65] I. D. Couzin, J. Krause, R. James, G. Ruxton, and N. Franks. Collective memory and spatial sorting in animal groups. *J. theor. Biol.*, 218:1–11, 2002.

[66] E. Cristiani, B. Piccoli, and A. Tosin. Modeling self-organization in pedestrians and animal groups from macroscopic and microscopic viewpoints. *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, pages 337–364, 2010.

[67] F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Trans. Automat. Control*, 52(5):852–862, 2007.

[68] F. Cucker and S. Smale. On the mathematics of emergence. *Jpn. J. Math.*, 2(1):197–227, 2007.

[69] F. T. Da Silva, H. M. B. F. Brito, and R. L. Pimentel. Modeling of crowd load in vertical direction using biodynamic model for pedestrians crossing footbridges. *Canadian Journal of Civil Engineering*, 40(12):1196–1204, 2013.

[70] S. D'Amico, J.-S. Ardaens, S. De Florio, and O. Montenbruck. Autonomous formation flying - Tandem-x, Prisma and beyond. In *Proceedings of the 5th International Workshop on Satellite Constellation & Formation Flying*, 2008.

[71] P. Degond, J.-G. Liu, S. Motsch, and V. Panferov. Hydrodynamic models of self-organized dynamics: derivation and existence theory. *Methods and Applications of Analysis*, 20:89–114, 2013.

[72] P. Degond and S. Motsch. Macroscopic limit of self-driven particles with orientation interaction. *C. R. Math. Acad. Sci. Paris*, 345(10):555–560, 2007.

[73] P. Degond and S. Motsch. Continuum limit of self-driven particles with orientation interaction. *Math. Models Methods Appl. Sci.*, 18(suppl.):1193–1215, 2008.

[74] P. Degond and S. Motsch. Continuum limit of self-driven particles with orientation interaction. *Math. Models Methods Appl. Sci.*, 18(suppl.):1193–1215, 2008.

[75] P. Degond and S. Motsch. A macroscopic model for a system of swarming agents using curvature control. *J. Stat. Phys.*, 143(4):685–714, 2011.

[76] A. Demeulemeester, C.-F. Hollemeersch, P. Mees, B. Pieters, P. Lambert, and R. Walle. Hybrid path planning for massive crowd simulation on the gpu. In J. Allbeck and P. Faloutsos, editors, *Motion in Games*, volume 7060 of *Lecture Notes in Computer Science*, pages 304–315. Springer Berlin Heidelberg, 2011.

[77] G. Dimarco, R. Caflisch, and L. Pareschi. Direct simulation Monte Carlo schemes for Coulomb interactions in plasmas. *Commun. Appl. Ind. Math.*, 1(1):72–91, 2010.

[78] G. Dimarco and L. Pareschi. Asymptotic preserving implicit-explicit Runge-Kutta methods for nonlinear kinetic equations. *SIAM J. Numer. Anal.*, 51(2):1064–1087, 2013.

[79] A. L. Dontchev, William, and W. Hager. The euler approximation in state constrained optimal control. *Mathematics of Computation*, 70:173–203, 1997.

[80] M. R. D'Orsogna, Y. Chuang, A. Bertozzi, and L. Chayes. Self-propelled particles with soft-core interactions: patterns, stability and collapse. *Phys. Rev. Lett.*, 96(104302), 2006.

[81] B. Düring, P. Markowich, J.-F. Pietschmann, and M.-T. Wolfram. Boltzmann and Fokker–Planck equations modelling opinion formation in the presence of strong leaders. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 465(2112):3687–3708, 2009.

[82] A. Dussutour, S. Nicolis, J.-L. Deneubourg, and V. Fourcassié. Collective decision in ants under crowded conditions. *Behavioral Ecology and Sociobiology*, 61(17-30), 2006.

[83] E. Fernández-Juricic, J. Erichsen, and A. Kacelnik. Visual perception and social foraging in birds. *Trends in Ecology & Evolution*, 19(1):25–31, 2004.

[84] E. Fernández-Juricic, S. Siller, and A. Kacelnik. Flock density, social foraging, and scanning: an experiment with starlings. *Behavioral Ecology*, 15(3):371, 2004.

[85] M. Fornasier, J. Haskovec, and G. Toscani. Fluid dynamic description of flocking via Povzner-Boltzmann equation. *Physica D*, 240:21–31, 2011.

[86] M. Fornasier, J. Haskovec, and J. Vybiral. Particle systems and kinetic equations modeling interacting agents in high dimension. *Multiscale Modeling & Simulation*, 9(4):1727–1764, 2011.

[87] M. Fornasier, B. Piccoli, and F. Rossi. Sparse mean-field optimal control. *preprint*, 2014.

[88] M. Fornasier and F. Solombrino. Mean-field optimal control. *preprint*, 2013.

[89] G. Furioli, A. Pulvirenti, E. Terraneo, and G. Toscani. The grazing collision limit of the inelastic Kac model around a Lévy-type equilibrium. *SIAM J. Math. Anal.*, 44(2):827–850, 2012.

[90] S. Galam, Y. Gefen, and Y. Shapir. Sociophysics: A new approach of sociological collective behavior. *J. Math. Sociology*, 9:1–13, 1982.

[91] V. Gazi and K. Passino. Stability analysis of swarms. *IEEE Trans. Auto. Control*, 48:692–697, 2003.

[92] M. Giles. Analysis of the accuracy of shock-capturing in the steady quasi 1d-Euler equations. *Int. J. Comput. Fluid Dynam.*, 5:247–258, 1996.

[93] M. Giles and N. A. Pierce. *Adjoint error correction for integral outputs*, volume 25 of *Lect. Notes Comput. Sci. Eng.*, pages 47–95. Springer Verlag, Berlin, 2003.

[94] M. Giles and E. Sueli. Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numerica*, 11:145–236, 2002.

[95] M. Giles and S. Ulbrich. Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws: Part 1:Linearized approximations and linearized output functional. *SIAM J. Numer. Anal.*, 48:882–904, 2010.

[96] M. Giles and S. Ulbrich. Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws: Part 2: Adjoint approximations and extensions. *SIAM J. Numer. Anal.*, 48:905–921, 2010.

[97] J. Gómez-Serrano, C. Graham, and J.-Y. Le Boudec. The bounded confidence model of opinion dynamics. *Math. Models Methods Appl. Sci.*, 22:1–46, 2012.

[98] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.

[99] S. Ha and J. Liu. A simple proof of the Cucker-Smale flocking dynamics and mean-field limit. *Commun. Math. Sci.*, 7(2):297–325, 2009.

[100] S. Y. Ha and E. Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *Kinet. Relat. Models*, 1(3):415–435, 2008.

[101] W. W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87:247–282, 1999.

[102] R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence, models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3):2, 2002.

[103] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Phys. Rev. E*, 51:4282–4286, May 1995.

[104] C. K. Hemelrijk and H. Hildenbrandt. Self-organized shape and frontal density of fish schools. *Ethology*, 114:245–254, 2008.

[105] C. K. Hemelrijk and H. Hildenbrandt. Schools of fish and flocks of birds: their shape and internal structure by self-organization. *Interface Focus*, 2(6):726–737, 2012.

[106] M. Herty and L. Pareschi. Fokker-planck asymptotics for traffic flow models. *Kinet. Relat. Models*, 3(1), 2010.

[107] M. Herty, L. Pareschi, and S. Steffensen. Implicit-explicit Runge-Kutta schemes for numerical discretization of optimal control problems. *SIAM J. Numer. Anal.*, 51(4):1875–1899, 2013.

[108] M. Herty and C. Ringhofer. Averaged kinetic models for flows on unstructured networks. *Kinet. Relat. Models*, 4(4):1081–1096, 2011.

[109] M. Herty and C. Ringhofer. Feedback controls for continuous priority models in supply chain management. *Comput. Methods Appl. Math.*, 11(2):206–213, 2011.

[110] M. Herty and A. N. Sandjo. On optimal treatment planning in radiotherapy governed by transport equations. *Math. Models Methods Appl. Sci.*, 21(2):345–359, 2011.

[111] M. Herty and V. Schleper. Time discretizations for numerical optimisation of hyperbolic problems. *Applied Mathematics and Computation*, 218(1):183 – 194, 2011.

[112] I. Higueras. Strong stability for additive Runge-Kutta methods. *SIAM J. Numer. Anal.*, 44(4):1735–1758 (electronic), 2006.

[113] H. Hildenbrandt, C. Carere, and C. Hemelrijk. Self-organized aerial displays of thousands of starlings: a model. *Behavioral Ecology*, 21(6):1349, 2010.

[114] A. Huth and C. Wissel. The simulation of fish schools in comparison with experimental data. *Ecol. Model.*, 75/76:135–145, 1994.

[115] C. A. Kennedy and M. H. Carpenter. Additive Runge–Kutta schemes for convectiondiffusion-reaction equations. *Appl. Numer. Math.*, 44(1-2):139–181, 2003.

[116] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 4, pages 1942–1948 vol.4, Nov 1995.

[117] A. L. Koch and D. White. The social lifestyle of myxobacteria. *BioEssays*, 20(12):1030–1038, 1998.

[118] T. Kolokolnikov, Y. Huang, and M. Pavlovski. Singular patterns for an aggregation model with a confining potential. *to appear in Physica D.*

[119] T. Kolokonikov, H. Sun, D. Uminsky, and A. Bertozzi. Stability of ring patterns arising from 2d particle interactions. *Physical Review E*, 84(1):015203, 2011.

[120] T. Krogstad. *Attitude synchronization in spacecraft formations: Theoretical and experimental results.* PhD Thesis, NTNU, Trondheim. 2009.

[121] M. Krstic, I. Kanellakopoulos, and P. Kokotovic. *Nonlinear and adaptive control design.* John Wiley and Sons Inc., New York, 1995.

[122] O. E. Lanford, III. On a derivation of the Boltzmann equation. In *International Conference on Dynamical Systems in Mathematical Physics (Rennes, 1975)*, pages 117–137. Astérisque, No. 40. Soc. Math. France, Paris, 1976.

[123] J. Lang and J. Verwer. W-methods in optimal control. *Numerische Mathematik*, 124(2):337–360, 2013.

[124] J. Lee, S. Cho, and R. Calvo. A fast algorithm for simulation of flocking behavior. In *Games Innovations Conference, 2009. ICE-GIC 2009. International IEEE Consumer Electronics Society's*, pages 186–190. IEEE, 2009.

[125] M. Lemou. Multipole expansions for the Fokker-Planck-Landau operator. *Numer. Math.*, 78(4):597–618, 1998.

[126] K. Lerman, A. Martinoli, and A. Galstyan. A review of probabilistic macroscopic models for swarm robotic systems. In E. Sahin and W. Spears, editors, *Swarm Robotics*, volume 3342 of *Lecture Notes in Computer Science*, pages 143–152. Springer Berlin Heidelberg, 2005.

[127] H. Levine, W.-J. Rappel, and I. Cohen. Self-organization in systems of self-propelled particles. *Phys. Rev. E*, 63:017101, Dec 2000.

[128] T. Liao, K. Socha, M. Montes de Oca, T. Stützle, and M. Dorigo. Ant colony optimization for mixed-variable optimization problems. *IEEE Transactions on Evolutionary Computation*, page in press, 2013.

[129] R. Lukeman, Y. Li, and L. Edelstein-Keshet. Inferring individual rules from collective behavior. *Proc. Natl. Acad. Sci. U.S.A.*, 107(28):12576–12580, 2010.

[130] T. Lux and M. Marchesi. Scaling and criticality in a stochastich multi-agent model of a financial market. *Nature*, 397:498–500, 1999.

[131] D. Maldarella and L. Pareschi. Kinetic models for socio–economic dynamics of speculative markets. *Physica A*, 391:715–730, 2012.

[132] D. Q. Mayne and H. Michalska. Receding horizon control of nonlinear systems. *IEEE Trans. Automat. Control*, 35(7):814–824, 1990.

[133] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert. Constrained model predictive control: stability and optimality. *Automatica J. IFAC*, 36(6):789–814, 2000.

[134] S. McNamara and W. Young. Kinetics of a one-dimensional granular medium in the quasi-elastic limit. *Phys. Fluids A*, 5:34–45, 1995.

[135] H. Michalska and D. Q. Mayne. Robust receding horizon control of constrained nonlinear systems. *IEEE Trans. Automat. Control*, 38(11):1623–1633, 1993.

[136] D. Morgan, S.-J. Chung, and F. Y. Hadaegh. Decentralized model predictive control of swarm of spacecraft using sequential convex programming. *Advances in the Astronautical Sciences Volume*, 148, 2103.

[137] N. Moshtagh, N. Michael, A. Jadbabaie, and K. Daniilidis. Vision-based, distributed control laws for motion coordination of nonholonomic robots. *Robotics, IEEE Transactions on*, 25(4):851–860, 2009.

[138] S. Motsch and E. Tadmor. A new model for self-organized dynamics and its flocking behavior. *Journal of Statistical Physics*, 144(5):923–947, 2011.

[139] S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *Preprint arXiv:1301.4123*, 2013.

[140] C. Mouhot and L. Pareschi. Fast algorithms for computing the Boltzmann collision operator. *Math. Comp.*, 75(256):1833–1852 (electronic), 2006.

[141] G. Naldi, L. Pareschi, and G. Toscani, editors. *Mathematical modeling of collective behavior in socio-economic and life sciences.* Modeling and Simulation in Science, Engineering and Technology. Birkhäuser Boston Inc., Boston, MA, 2010.

[142] A. Okubo. Dynamical aspects of animal grouping: swarms, schools, flocks, and herds. *Advances in Biophysics*, 22:1–94, 1986.

[143] L. Pareschi and G. Russo. An introduction to Monte Carlo methods for the Boltzmann equation. In *CEMRACS 1999 (Orsay)*, volume 10 of *ESAIM Proc.*, pages 35–76. Soc. Math. Appl. Indust., Paris, 1999.

[144] L. Pareschi and G. Russo. Recent trends in numerical analysis. chapter Implicit-explicit Runge-Kutta Schemes for Stiff Systems of Differential Equations, pages 269–288. Nova Science Publishers, Inc., Commack, NY, USA, 2000.

[145] L. Pareschi and G. Russo. Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *Journal of Scientific Computing*, 25(1):129–155, 2005.

[146] L. Pareschi and G. Toscani. *Interacting multi-agent systems. Kinetic equations & Monte Carlo methods.* Oxford University Press, USA, 2013.

[147] J. Parrish and L. Edelstein-Keshet. Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science*, 284(5411):99 –101, 1999.

[148] L. Perea, G. Gomez, and P. Elosegui. Extension of the Cucker–Smale control law to space flight formations. *AIAA Journal of Guidance, Control, and Dynamics*, 32:527–537, 2009.

[149] A. J. Povzner. On the Boltzmann equation in the kinetic theory of gases. *Mat. Sb. (N.S.)*, 58 (100):65–86, 1962.

[150] C. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *ACM SIGGRAPH Computer Graphics*, volume 21, pages 25–34. ACM, 1987.

[151] J. Sanz-Serna. Runge-Kutta schemes for Hamiltonian systems. *BIT Numerical Mathematics*, 28(4):877–883, 1988.

[152] J. M. Sanz-Serna and L. Abia. Order conditions for canonical runge-kutta schemes. *SIAM Journal on Numerical Analysis*, 28(4):pp. 1081–1096, 1991.

[153] E. D. Sontag. *Mathematical control theory*, volume 6 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 1998. Deterministic finite-dimensional systems.

[154] K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community, 2000.

[155] G. Toscani. Kinetic models of opinion formation. *Commun. Math. Sci.*, 4(3):481–496, 2006.

[156] F. Tröltzsch and D. Wachsmuth. On convergence of a receding horizon method for parabolic boundary control. *Optim. Methods Softw.*, 19(2):201–216, 2004.

[157] S. Ulbrich. Adjoint-based derivative computations for the optimal control of discontinuous solutions of hyperbolic conservation laws. *System Control Letters*, 48:313–328, 2003.

[158] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.*, 75(6):1226–1229, 1995.

[159] T. Vicsek and A. Zafeiris. Collective motion. *Physics Reports*, 517(3–4):71 – 140, 2012. Collective motion.

[160] C. Villani. On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations. *Arch. Rational Mech. Anal.*, 143(3):273–307, 1998.

[161] J. von Brecht, D. Uminsky, T. Kolokolnikov, and A. Bertozzi. Predicting pattern formation in particle interactions. *Math. Mod. Meth. Appl. Sci.*, 22:1140002, 2012.

[162] A. Walther. Automatic differentiation of explicit Runge-Kutta methods for optimal control. *Computational Optimization and Applications*, 36(1):83–108, 2007.

[163] X. Wang, X. Jin, Z. Deng, and L. Zhou. Inherent noise-aware insect swarm simulation. *Computer Graphics Forum*, 2014.

[164] C. A. Yates, R. Erban, C. Escudero, I. D. Couzin, J. Buhl, I. G. Kevrekidis, P. K. Maini, and D. J. T. Sumpter. Inherent noise can facilitate coherence in collective swarm motion. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5464–5469, 2010.

[165] Q. Zhang, M. Liu, J. Liu, and G. Zhao. Modification of evacuation time computational model for stadium crowd risk analysis. *Process Safety and Environmental Protection*, 85(6):541 – 548, 2007.

Acknowledgments

I would like to express my special appreciation and thanks to my advisor Prof. Lorenzo Pareschi, for his good advice and friendship, for all the sound and fruitful discussions, this work wouldn't have been possible without his help.

I would also like to thank Prof. José A. Carrillo for his valuable advices and exchanges of ideas, his way of enjoying research has been a great inspiration to me. Moreover I thank him for his kind hospitality during my staying in Barcelona and London.

Further I thank Prof. Michael Herty, for his precious help and nice hospitality, coming to Aachen has been always a pleasure, thank to him and his research team.

A big thank to all my colleagues and friends I met around the world, especially here in Ferrara. I really enjoyed the time we spent together, with you it's kind of feeling at home.

A huge thank to my friends in Verona, because the time we spend together is always a time for sharing dreams, projects and also simple genuine fun.

The most important thank goes to my family for supporting me throughout my life, for all the teachings I received and because coming back home it gives me always that warm feeling.

Finally, to all the people I've walked this path with:

> *"No one can pass through life, any more than he can pass through a bit of country, without leaving tracks behind, and those tracks may often be helpful to those coming after him in finding their way."*
>
> Baden Powell