

Modeling and Simulation in Science,
Engineering and Technology

Nicola Bellomo
Pierre Degond
Eitan Tadmor
Editors

Active Particles, Volume 1

Advances in Theory, Models, and
Applications



Birkhäuser



Modeling and Simulation in Science, Engineering and Technology

Series editors

Nicola Bellomo
Department of Mathematics
Faculty of Sciences
King Abdulaziz University
Jeddah, Saudi Arabia

Tayfun E. Tezduyar
Department of Mechanical Engineering
Rice University
Houston, TX, USA

Editorial Advisory Board

K. Aoki
Kyoto University
Kyoto, Japan

P. Koumoutsakos
Computational Science & Engineering
Laboratory
ETH Zürich
Zürich, Switzerland

K.J. Bathe
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA, USA

H.G. Othmer
Department of Mathematics
University of Minnesota
Minneapolis, MN, USA

Y. Bazilevs
Department of Structural Engineering
University of California, San Diego
La Jolla, CA, USA

K.R. Rajagopal
Department of Mechanical Engineering
Texas A&M University
College Station, TX, USA

M. Chaplain
Division of Mathematics
University of Dundee
Dundee, Scotland, UK

K. Takizawa
Department of Modern Mechanical
Engineering
Waseda University
Tokyo, Japan

P. Degond
Department of Mathematics
Imperial College London
London, UK

Y. Tao
Dong Hua University
Shanghai
China

A. Deutsch
Center for Information Services
and High-Performance Computing
Technische Universität Dresden
Dresden, Germany

M.A. Herrero
Departamento de Matematica Aplicada
Universidad Complutense de Madrid
Madrid, Spain

More information about this series at <http://www.springer.com/series/4960>

Nicola Bellomo · Pierre Degond
Eitan Tadmor
Editors

Active Particles, Volume 1

Advances in Theory, Models,
and Applications



Editors

Nicola Bellomo

Department of Mathematical Sciences
King Abdulaziz University
Jeddah
Saudi Arabia

Eitan Tadmor

CSCAMM
University of Maryland
College Park
USA

Pierre Degond

Department of Mathematics
Imperial College London
London
UK

ISSN 2164-3679

ISSN 2164-3725 (electronic)

Modeling and Simulation in Science, Engineering and Technology

ISBN 978-3-319-49994-9

ISBN 978-3-319-49996-3 (eBook)

DOI 10.1007/978-3-319-49996-3

Library of Congress Control Number: 2016961339

Mathematics Subject Classification (2010): 35K55, 35Q92, 35Q70, 37N40, 60H10, 49J15, 74A25, 70F45, 76N10, 82D99, 91A26, 92D25, 91C20, 82B21, 35Q91, 49J45

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This book is published under the trade name Birkhäuser

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland
(www.birkhauser-science.com)

Preface

This edited book collects ten surveys on modeling, simulation, and applications of active matter by various types of methods spanning from mathematical kinetic theory to nonequilibrium statistical mechanics. The contributions develop a variety of different viewpoints, such as individual-based models, evolutive games, Brownian particles, kinetic theory, and continuum theories, as well as a combination of these, such as kinetic theory and evolutive games or individual-based and continuum models. The authors of each chapter have been selected among the most active mathematicians operating on the aforementioned challenging field. The contents provide a survey of recent results of the various teams and look ahead to research perspectives. Hence, this book is just timely to provide the scientific community with an up-to-date overview of current research conducted by leading experts in this field.

Indeed, this book covers a broad range of applications, such as biological network formation and network theory in the chapters “[Continuum Modeling of Biological Network Formation](#)” and “[Interaction Network, State Space, and Control in Social Dynamics](#)”; opinion formation and social systems in the chapters “[Recent Advances in Opinion Modeling: Control and Social Influence](#),” “[Interaction Network, State Space, and Control in Social Dynamics](#),” “[Sparse Control of Multiagent Systems](#),” and “[A Review on Attractive–Repulsive Hydrodynamics for Consensus in Collective Behavior](#); control theory of sparse systems in the chapters “[Recent Advances in Opinion Modeling: Control and Social Influence](#),” “[Interaction Network, State Space and Control in Social Dynamics](#),” and “[Sparse Control of Multiagent Systems](#); theory and applications of mean field games in the chapter “[Variational Mean Field Games](#);” population learning in the chapter “[Sparse Control of Multiagent Systems](#);” dynamics of flocking systems in the chapter “[Emergent Dynamics of the Cucker–Smale Flocking Model and Its Variants](#);” vehicular traffic in the chapter “[Follow-the-Leader Approximations of Macroscopic Models for Vehicular and Pedestrian Flows](#);” and stochastic particles and mean field approximation in the chapter “[Mean Field Limit for Stochastic Particle Systems](#). ”

Different mathematical tools have been used from methods of generalized kinetic theory and statistical dynamics to stochastic evolutive games, mean field games, and stochastic differential systems. Control theory, flocking analysis, and network theory contribute and enrich the application of the aforementioned mathematical tools.

The variety of applications and the interdisciplinary use of different mathematical tools witness the interest of applied mathematicians toward modeling, qualitative analysis, and computing of large systems of active particles viewed as living, hence complex, systems. This new frontier of science offers to applied mathematicians a broad variety of new challenging problems.

The research activity in the field meets an equally productive scientific environment. In particular, we mention the Ki-Net, an NSF Research Network focused on “Kinetic description of emerging challenges in multiscale problems of natural sciences” (www.ki-net.umd.edu). The Ki-Net, through its main three hubs in the University of Maryland, University of Wisconsin, and UT Austin and an interlinked network of 20+ nodes, fosters a series of activities with a main intellectual focus on development, analysis, computation, and application of quantum dynamics, network dynamics, and kinetic models of biological processes. As such, Ki-Net is a primary outlet for presentation of recent activities in the above areas of active matter. Indeed, many of the authors in this special volume were involved in Ki-Net activities. We mention in this context the recent examples of Ki-Net workshops on modeling and control in social dynamics (October 2014), on groups and interactions in data, networks and biology (May 2015), and on collective dynamics in biological and social systems (November 2015). A complete list of activities can be found at www.ki-net.umd.edu/content/activities, and we use this opportunity to acknowledge the NSF support of Ki-Net Grant #1107444 for funding these activities.

Jeddah, Saudi Arabia
London, UK
College Park, USA

Nicola Bellomo
Pierre Degond
Eitan Tadmor

Contents

Continuum Modeling of Biological Network Formation	1
Giacomo Albi, Martin Burger, Jan Haskovec, Peter Markowich and Matthias Schlottbom	
Recent Advances in Opinion Modeling: Control and Social Influence	49
Giacomo Albi, Lorenzo Pareschi, Giuseppe Toscani and Mattia Zanella	
Interaction Network, State Space, and Control in Social Dynamics	99
Aylin Aydoğdu, Marco Caponigro, Sean McQuade, Benedetto Piccoli, Nastassia Pouradier Duteil, Francesco Rossi and Emmanuel Trélat	
Variational Mean Field Games	141
Jean-David Benamou, Guillaume Carlier and Filippo Santambrogio	
Sparse Control of Multiagent Systems	173
Mattia Bongini and Massimo Fornasier	
A Kinetic Theory Approach to the Modeling of Complex Living Systems	229
Diletta Burini, Livio Gibelli and Nisrine Outada	
A Review on Attractive–Repulsive Hydrodynamics for Consensus in Collective Behavior	259
José A. Carrillo, Young-Pil Choi and Sergio P. Perez	
Emergent Dynamics of the Cucker–Smale Flocking Model and Its Variants	299
Young-Pil Choi, Seung-Yeal Ha and Zhuchun Li	
Follow-the-Leader Approximations of Macroscopic Models for Vehicular and Pedestrian Flows	333
M. Di Francesco, S. Fagioli, M.D. Rosini and G. Russo	
Mean Field Limit for Stochastic Particle Systems	379
Pierre-Emmanuel Jabin and Zhenfu Wang	

Contributors

Giacomo Albi Applied Numerical Analysis, Technical University of Munich, Garching bei München, Germany

Aylin Aydoğdu Rutgers University, Camden, NJ, USA

Jean-David Benamou INRIA-MOKAPLAN, Paris, France

Mattia Bongini Technische Universität München, Garching, Germany

Martin Burger Institute for Computational and Applied Mathematics, University of Münster, Münster, Germany

Diletta Burini Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy

Marco Caponigro Conservatoire National des Arts et Métiers, Paris, France

Guillaume Carlier Ceremade Univ. Paris Dauphine and INRIA-MOKAPLAN, Paris, France

José A. Carrillo Department of Mathematics, Imperial College London, South Kensington, UK

Young-Pil Choi Fakultät für Mathematik, Technische Universität München, Garching bei München, Germany

Nastassia Pouradier Duteil Rutgers University, Camden, NJ, USA

S. Fagioli DISIM, Università degli Studi dell’Aquila, L’Aquila (AQ), Italy

Massimo Fornasier Technische Universität München, Garching, Germany

M. Di Francesco DISIM, Università degli Studi dell’Aquila, L’Aquila (AQ), Italy

Livio Gibelli Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy

Seung-Yeal Ha Department of Mathematical Sciences and Research Institute of Mathematics, Seoul National University, Seoul, Republic of Korea

Jan Haskovec Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

Pierre-Emmanuel Jabin CSCAMM and Department of Mathematics, University of Maryland, College Park, MD, USA

Zhuchun Li Department of Mathematics, Harbin Institute of Technology, Harbin, People's Republic of China

Peter Markowich Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

Sean McQuade Rutgers University, Camden, NJ, USA

Nisrine Outada Faculté des Sciences Semlalia, Marrakesh, Morocco; Laboratoire Jacques Louis-Lions, Université Pierre et Marie Curie, Paris, France

Lorenzo Pareschi University of Ferrara, Ferrara, Italy

Sergio P. Perez ETSIAE, Technical University of Madrid, Madrid, Spain

Benedetto Piccoli Rutgers University, Camden, NJ, USA

M.D. Rosini Instytut Matematyki, Uniwersytet Marii Curie-Skłodowskiej, Lublin, Poland

Francesco Rossi Aix Marseille Université, CNRS, ENSAM, Université de Toulon, Marseille, France

G. Russo Dipartimento di Matematica ed Informatica, Università di Catania, Catania, Italy

Filippo Santambrogio Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay Cedex, France

Matthias Schlottbom Multiscale Modeling and Simulation, University of Twente, AE Enschede, The Netherlands

Giuseppe Toscani University of Pavia, Pavia, Italy

Emmanuel Trélat Sorbonne Universités, Paris, France; Laboratoire Jacques-Louis Lions, Institut Universitaire de France, Paris, France

Zhenfu Wang CSCAMM and Department of Mathematics, University of Maryland, College Park, MD, USA

Mattia Zanella University of Ferrara, Ferrara, Italy

Continuum Modeling of Biological Network Formation

Giacomo Albi, Martin Burger, Jan Haskovec, Peter Markowich and
Matthias Schlottbom

Abstract We present an overview of recent analytical and numerical results for the elliptic–parabolic system of partial differential equations proposed by Hu and Cai, which models the formation of biological transportation networks. The model describes the pressure field using a Darcy type equation and the dynamics of the conductance network under pressure force effects. Randomness in the material structure is represented by a linear diffusion term and conductance relaxation by an algebraic decay term. We first introduce micro- and mesoscopic models and show how they are connected to the macroscopic PDE system. Then, we provide an overview of analytical results for the PDE model, focusing mainly on the existence of weak and mild solutions and analysis of the steady states. The analytical part is complemented by extensive numerical simulations. We propose a discretization based on finite elements and study the qualitative properties of network structures for various parameter values.

G. Albi

Applied Numerical Analysis, Technical University of Munich, Boltzmannstr. 3,
85478 Garching bei München, Germany
e-mail: giacomo.albi@ma.tum.de

M. Burger

Institute for Computational and Applied Mathematics, University of Münster,
Einsteinstr. 62, 48149 Münster, Germany
e-mail: martin.burger@wwu.de

J. Haskovec (✉) · P. Markowich

Mathematical and Computer Sciences and Engineering Division, King Abdullah
University of Science and Technology, Thuwal 23955, Kingdom of Saudi Arabia
e-mail: jan.haskovec@kaust.edu.sa

P. Markowich

e-mail: peter.markowich@kaust.edu.sa

M. Schlottbom

Multiscale Modeling and Simulation, University of Twente,
P.O. Box 217, NL-7500 AE Enschede, The Netherlands
e-mail: m.schlottbom@utwente.nl

1 Introduction

A *transportation network* is a realization of a spatial structure which permits flow of some commodity. Transportation networks are ubiquitous in both social and biological systems and play a vital role in virtually all aspects of everyday life. Robust network performance involves a complex trade-off involving cost, transportation efficiency, and fault tolerance [4]. Biological transportation networks develop without centralized control [44] and have been fine-tuned by many cycles of evolutionary selection pressure. They can therefore be considered optimal solutions of the underlying transportation problems at which cost, efficiency, and resilience are appropriately balanced [11]. Moreover, they typically afford great benefit to biological systems; for instance, the elastic property of a leaf is intimately related to the structure of its leaf venation [30].

This chapter is devoted to mathematical modeling of *biological transportation networks in porous media*, for instance, leaf venation in plants [12, 31], mammalian circulatory systems that convey nutrients to the body through blood circulation [42, 45], or neural networks that transport electric charge [16, 33]. In particular, we focus on the partial differential equation (PDE)-based model recently introduced in [24] and studied in the series of papers [1, 20, 22],

$$-\nabla \cdot [(rI + m \otimes m)\nabla p] = S, \quad (1)$$

$$\frac{\partial m}{\partial t} - D^2 \Delta m - c^2(m \cdot \nabla p)\nabla p + |m|^{2(\gamma-1)}m = 0, \quad (2)$$

where $p = p(t, x) \in \mathbb{R}$ is the scalar pressure of the fluid transported within the network and $m = m(t, x) \in \mathbb{R}^d$ is the vector-valued conductance, with $d \in \{1, 2, 3\}$ the space dimension. The parameters are $D^2 \geq 0$ (diffusivity), $c^2 > 0$ (activation parameter), and $\gamma \in \mathbb{R}$. The scalar function $r = r(x) \geq r_0 > 0$ describes the isotropic background permeability of the medium. The term $S = S(x)$ models sources and sinks that we for simplicity assume time-independent.

The system (1), (2) is posed on a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, with smooth boundary $\partial\Omega$. We prescribe the homogeneous Dirichlet boundary conditions on $\partial\Omega$ for m and p ,

$$m(t, x) = 0, \quad p(t, x) = 0 \quad \text{for } x \in \partial\Omega, \quad t \geq 0, \quad (3)$$

and the initial condition for m ,

$$m(t = 0, x) = m^I(x) \quad \text{for } x \in \Omega. \quad (4)$$

In Section 2.2, we give a short overview of discrete network models, focusing mainly on the recent dynamic model of Hu and Cai [25]. In Section 2.2, we establish a connection between their model and the PDE system (1)–(2), and in Section 3, we provide an overview of results concerning its mathematical analysis. We discuss

numerical schemes for its discretization in Section 4 and present results of systematic numerical simulations for the two-dimensional problem in Section 5.

2 Modeling

In the following, we discuss different modeling approaches: We start with a derivation of the continuum model based on macroscopic physical laws and then discuss microscopic models proposed in the literature. Subsequently, we connect them by some formal arguments as well as a novel mesoscopic model.

2.1 Macroscopic Model

Let the network domain $\Omega \subset \mathbb{R}^d$ be occupied by a porous medium in which a fluid moves with velocity $v = v(t, x) \in \mathbb{R}^d$. For the sake of simplicity, we assume that v is a smooth function. We denote by $\rho = \rho(t, x)$ the mass density of the fluid, and assuming that the fluid is injected into or expelled from the medium at rate $S = S(x)$, the density satisfies the mass continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = \rho S.$$

Moreover, assuming the quasi-incompressibility of the fluid, i.e., constant fluid density along particle trajectories, we have

$$\frac{D\rho}{Dt} = \frac{\partial \rho}{\partial t} + v \cdot \nabla \rho = 0.$$

Combined with the previous equation, we obtain

$$\nabla \cdot v = S. \quad (5)$$

Finally, we assume the validity of Darcy's law for slow flow in porous media [46], therefore the velocity v is given by

$$v = -\mathbb{P}[m] \nabla p,$$

where $\mathbb{P}[m]$ is the permeability tensor, depending on the network conductance vector $m \in \mathbb{R}^d$, and p the fluid pressure. Taking $\mathbb{P}[m]$ of the form

$$\mathbb{P}[m] = r I + m \otimes m, \quad (6)$$

yields the Poisson equation

$$-\nabla \cdot ((rI + m \otimes m)\nabla p) = S, \quad (7)$$

where I is the identity matrix, and the scalar function $r = r(x) \geq r_0 > 0$ describes the isotropic background permeability of the medium. Then (7) subject to the boundary datum (3) has a unique weak solution $p = p[m]$ for each $m \in L^\infty(\Omega)$. Note that $\mathbb{P}[m]$ has the principal directions (eigenvectors) $m/|m|$ and, resp., m^\perp with eigenvalues $r(x) + |m|^2$ and, resp., $r(x)$. This means that the flow "feels" only the background permeability $r = r(x)$ in the directions orthogonal to m , while the principal permeability is increased by $|m|^2$ in the direction of the conductance vector m . A simple calculation reveals that (1), (2) with $D = 0$ is the formal L^2 -gradient flow of the energy

$$E_1[m] := \frac{1}{2} \int c^2 (\nabla p[m] \cdot \mathbb{P}[m] \nabla p[m]) + \frac{|m|^{2\gamma}}{\gamma} dx.$$

To model the propagation of the network structure, we add a diffusion term to obtain (2). Denoting by $D^2 \geq 0$ the diffusivity constant, the modified energy reads as

$$\mathcal{E}[m] := \frac{1}{2} \int D^2 |\nabla m|^2 + c^2 (\nabla p[m] \cdot \mathbb{P}[m] \nabla p[m]) + \frac{|m|^{2\gamma}}{\gamma} dx. \quad (8)$$

Denoting $p = p[m]$, the corresponding formal L^2 -gradient flow of $\mathcal{E}[m]$, constrained by (7), is given by

$$\partial_t m = D^2 \Delta m + c^2 (m \cdot \nabla p) \nabla p - |m|^{2(\gamma-1)} m. \quad (9)$$

The parabolic reaction–diffusion equation (9) governs the evolution of the network conductance. The first term of the right-hand side $D^2 \Delta m$ represents random effects (Brownian process) in the network structure. The term $c^2 (m \cdot \nabla p) \nabla p$ is called the *activation term* and represents a driving force in the direction of the pressure gradient. The last term $-|m|^{2(\gamma-1)} m$ is the algebraic relaxation term, representing the functional derivative of the metabolic cost of maintaining the network.

2.2 Microscopic Models

Traditionally, models of biological transportation networks were based on *discrete* frameworks, in particular mathematical graph theory and discrete energy optimization, where the energy consumption within the networks is minimized under the constraint of constant total material cost [3, 5, 14].

One can classify the discrete mathematical models of transportation networks into *static* or *dynamic*, the latter ones accounting for adaptation of networks to fluctuations in the flow. For blood circulation systems, it is well known that throughout the life of humans and animals, blood vessel systems are continuously adapting their structures to meet the changing metabolic demand of the tissue. In particular, it has been observed in experiments that blood vessels can sense the wall shear stress [38] and adapt their diameters according to it [26]. Consequently, for biological applications it is necessary to employ the *dynamic* class of models. One such model was introduced in [28], and it was shown that the optimal blood vessel network can be achieved through adaptation of the vessel radius. It has been found that the adaptation to wall shear stress can lead to topological changes in the vessel network structure, and this adaptation of the local properties of the network, in particular the vessel radius, is related to the optimization of the global energy consumption of the network. An adaptive model was also introduced in the study of the networks formed by slime mold *Physarum polycephalum* [44], where networks at the steady states of adaptation are compared with networks with optimal total edge length.

One of the main research questions is what are the structural and topological properties of the optimal networks, in particular, existence of loops. In general, biological transportation networks do contain many loops [11, 30, 35]; for instance, the vasculature of two-dimensional animal tissues such as the retina and dicotyledon leaf venation are two of many natural networks which contain loops [29]. Animals and plants benefit from loop structures in many ways. For example, loops are important in mitigating damages of networks [29] and optimizing energy consumption with fluctuating flow distributions [11].

Central to this contribution is the new approach to *dynamic* modeling of transportation networks recently introduced by Hu and Cai [25]. Contrary to the global effect of optimization, they propose a purely local dynamic adaptation model that is based on mechanical laws. Their adaptation dynamics responds only to local information and can naturally incorporate fluctuations in the flow. Let us denote the finite, a priori given set of unoriented edges (or vessels) \mathcal{I} and the set of vertices \mathcal{V} . We assume that any pair of vertices is connected by at most one edge and that the corresponding graph $(\mathcal{V}, \mathcal{I})$ is connected. Let us emphasize that by fixing $(\mathcal{V}, \mathcal{I})$, the possible set of directions in the network is also fixed.

We will denote Q_i, L_i and, resp., C_i the flow, the length and, resp., the conductivity of the vessel $i \in \mathcal{I}$. Note that the conductivities C_i are intensive quantities (bulk property), while the conductances C_i/L_i are extensive (depending on the system size), see, e.g., [23]. Due to typically low Reynolds number of the flow, the flow rate in a vessel is proportional to the conductance and pressure drop (ΔP) between the two ends of the vessel $i \in \mathcal{I}$, $Q_i = C_i \frac{(\Delta P)_i}{L_i}$. Conservation of mass is expressed in terms of the Kirchhoff law,

$$\sum_{i \in N(j)} (P_j - P_{k(i,j)}) \frac{C_i}{L_i^2} = S_j \quad \text{for all } j \in \mathcal{V}, \quad (10)$$

Table 1 Notation

Variable	Meaning	Related to
Q_i	flow through an edge	edge $i \in \mathcal{I}$
L_i	length of an edge	edge $i \in \mathcal{I}$
C_i	conductivity	edge $i \in \mathcal{I}$
C_i/L_i	conductance	edge $i \in \mathcal{I}$
P_j	pressure	vertex $j \in \mathcal{V}$
S_j	intensity of source/sink	vertex $j \in \mathcal{V}$

where $N(j)$ denotes the set of edges adjacent to vertex j and $k(i, j) \in \mathcal{V}$ is the other vertex adjacent to edge i . In other words, edge i connects the vertices j and $k(i, j)$, and $P_j - P_{k(i, j)} = (\Delta P)_i$. S_j is the prescribed strength of flow source at node j , and we assume the conservation of total mass

$$\sum_{j \in \mathcal{V}} S_j = 0.$$

An overview of the notation is provided in Table 1. Note that for any given vector of conductivities $\mathbf{C} := (C_i)_{i \in \mathcal{I}}$, (10) represents a linear system of equations for the vector $(P_j)_{j \in \mathcal{V}}$. The system has a solution, unique up to an additive constant, if and only if the graph with edge weights given by \mathbf{C} is connected [18], where only edges with positive conductivities are taken into account (i.e., edges with zero conductivity are discarded).

Assuming that the material cost for an edge $i \in \mathcal{I}$ of the network is proportional to a power C_i^γ of its conductivity C_i , Hu and Cai consider the energy consumption function of the form

$$E_{\text{disc}}[\mathbf{C}] := \sum_{i \in \mathcal{I}} \left(\frac{Q_i(\mathbf{C})^2}{C_i} + \nu C_i^\gamma \right) L_i, \quad (11)$$

where $\nu > 0$ is a metabolic coefficient. The energy is constrained by the Kirchhoff law (10), i.e., $Q_i(\mathbf{C})$ is given by $Q_i(\mathbf{C}) = C_i \frac{(\Delta P)_i}{L_i}$ with the pressure drop $(\Delta P)_i$ being determined by (10) with conductivities \mathbf{C} . The first part of the energy consumption (11) is the kinetic energy (pumping power) of the material flow through the vessels. The second part represents the metabolic cost of maintaining the network. For instance, the metabolic cost for a blood vessel is proportional to its cross-sectional area [34]. Modeling blood flow by Hagen–Poiseuille’s law, the conductivity C_i is proportional to the square of the luminal diameter of the blood vessel. This implies $\gamma = 1/2$ for blood vessel systems. For leaf venations, the material cost may also be proportional to the number of small tubes, which is proportional to C_i , and the metabolic cost may be due to the effective loss of the photosynthetic power at the

area of the venation cells, which is proportional to $C_i^{1/\gamma}$. Consequently, the effective value of γ typically used in models of leaf venation lies between $1/2$ and 1 , [25].

Hu and Cai [25] consider the gradient flow of the energy (11) constrained by the Kirchhoff law (10), which is given by the ODE system

$$\frac{dC_i}{dt} = \left(\frac{\mathcal{Q}_i(\mathbf{C})^2}{C_i^2} - v\gamma C_i^{\gamma-1} \right) L_i, \quad (12)$$

noting that (10) implies $\frac{\partial \mathcal{Q}_i(\mathbf{C})}{\partial C_i} = 0$. They have shown that the optimal networks generated by (12) exhibit a phase transition at $\gamma = 1$, with a “uniform sheet” (the network is tiled with loops) for $\gamma > 1$ and a “loopless tree” for $\gamma < 1$.

2.2.1 Network Adaptation in the Hu-Cai Model

In what follows we show the phase transition behavior of the Hu-Cai model with respect to the parameter γ . We proceed by solving numerically a constrained energy minimization problem. We consider a planar graph $\mathcal{G} = (\mathcal{V}, \mathcal{I})$, such that the set of vertices and edges defines a diamond-shaped geometry contained in the domain $\Omega = (0, 2) \times (-1.5, 0.5)$, with $|\mathcal{V}| = 438$, $|\mathcal{I}| = 1233$, see Figure 1. We assume the source S to be positive on the subset of vertices $\Sigma^+ \subset \{(x, y) \in \Omega \mid x \leq 0.1\}$, and such that $S_j = \sigma_j^+ > 0$, with $\sigma_j^+ = 10^4 \exp(-(50x_j^2 + 10(y_j + 0.5)^4))$ for $j \in \Sigma^+$ and (x_j, y_j) the position of vertex j . On the complement of Σ^+ , $\Sigma^- \equiv \mathcal{V} \setminus \Sigma^+$, the source is constant and negative, $S_j = \sigma^- < 0$, and defined so that the total mass is conserved, i.e., $\sigma^- := -\sum_{k \in \Sigma^+} \sigma_k^+ / |\Sigma^-|$. Finally, we assume the initial conductivities \mathbf{C}^0 to be constant on \mathcal{I} and equal to $C_i = 10^4$ for every $i \in \mathcal{I}$.

In order to let the system adapt to an optimal network structure, we proceed solving iteratively the following steps:

Initialization. For every $i \in \mathcal{I}$, compute L_i and set the following parameters $\tau = 0.025$, $v = 1/\gamma$, $tol = 10^{-6}$.

1. Step: Pressure. Given the conductivities \mathbf{C}^0 , compute the pressure $\mathbf{P} = (P_j)_{j \in \mathcal{V}}$, solving the Kirchhoff law (10) by a least square minimization,

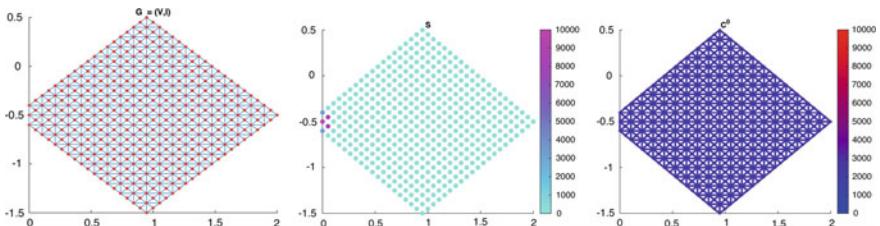


Fig. 1 Microscopic setting: Structure of the graph $G = (V, I)$, left, representation of the flow S on the vertices \mathcal{V} (middle). The initial conductivity is defined constant \mathbf{C}^0 on the edges \mathcal{I} (right).

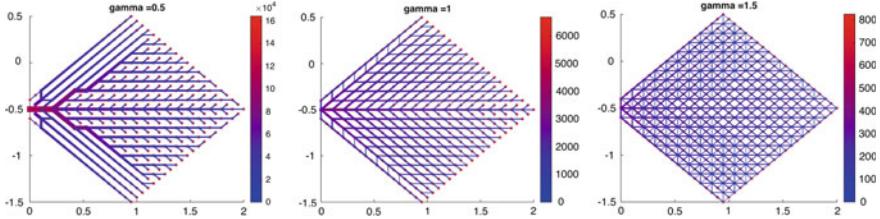


Fig. 2 Phase transition for γ : “loopless” tree with $\gamma = 0.5$ (left), network with “loops” for $\gamma = 1.5$ (right), and for $\gamma = 1$ an in-between state (middle). The edges’ width is proportional to $C_i^{1/5}$.

$$\min_{\mathbf{P}} \|\mathbb{H}[\mathbf{C}^0]\mathbf{P} - S\|_2, \quad (13)$$

where \mathbb{H} is the Laplacian matrix, defined as $(\mathbb{H}[\mathbf{C}])_{k(i,j),j} = -C_i/L_i^2$ for every $i \in N(j)$ and $(\mathbb{H}[\mathbf{C}])_{jj} = \sum_{i \in N(j)} (C_i/L_i^2)$.

2. Step: Conductivity. Given the pressure \mathbf{P} and conductivities \mathbf{C}^0 , find a minimizer \mathbf{C} of the following regularized version of the energy (11),

$$E_{\text{disc}}^\tau[\mathbf{C}] = \frac{\|\mathbf{C} - \mathbf{C}^0\|_2^2}{2\tau} + \sum_{i \in \mathcal{I}} \left(\left(\frac{\mathcal{Q}_i(\mathbf{C})^2}{C_i} + v(C_i)^\gamma \right) L_i \right), \quad (14)$$

where τ represents a regularization parameter. Note that for $\tau \rightarrow 0$, this minimization problem, constrained by the Kirchhoff’s law, is equivalent to solving (12), and for $\tau > 0$ amounts to an implicit Euler step of (12).

3. Step: Check energy decrease. Until $|E_{\text{disc}}^\tau[\mathbf{C}] - E_{\text{disc}}^\tau[\mathbf{C}^0]| > tol$, set $\mathbf{C}^0 = \mathbf{C}$, and go to step 1.

We report in Figure 2 the solutions obtained through this algorithmic procedure varying γ . The behavior of the solution shows the characteristic “loops” for $\gamma = 1.5$, “loopless” case for $\gamma = 0.5$ and an in-between state for $\gamma = 1$. We refer to [25, 29] for a phase transition study of this type of models. The numerical procedure described above is implemented in MATLAB, and to perform the minimization step (14) we rely on interior point method implemented in the CVX package, [17].

Remark 1 An alternative numerical approach would consist in solving directly the ODE system (12), coupled with the Kirchhoff law, as it has been done in [25]. However, for small values of C_i , the problem becomes stiff, therefore numerical strategies should assure the decay of the energy, e.g., by means of an adapting time step control. The minimization of (14) at Step 2 prevents this issue, since it works as an implicit solver. We discuss further the numerical treatment of this type of models in Sections 4.1 and 4.2.

2.3 Connection Between Micro- and Macroscopic Models

We shall now establish a formal connection between the discrete ODE system (12) and the continuum PDE description (1)–(2). On a bounded domain $\Omega \subset \mathbb{R}^d$, we consider the vector field $m = m(t, x)$ such that

$$\mathbb{P}[m] = m \otimes m$$

is the permeability tensor. Since in the discrete model we did not assume any background conductivity, we set $r = 0$ for the moment. Note that $\mathbb{P}[m]$ has the eigenvalues $|m|^2$ with eigenvector m , and 0 with eigenvectors orthogonal to m . Thus, it represents conduction along the direction m with conductivity $|m|^2$, while there is no conduction in directions perpendicular to m . Consequently, we locally identify the network conductivity with the principal eigenvalue of \mathbb{P} , i.e., $C_i \simeq M^2 := |m|^2$. This motivates us to introduce the variable transform $C_i := M_i^2$ and rewrite the discrete energy (11) in terms of $\mathbf{M} := (M_i)_{i \in \mathcal{I}}$ as

$$\bar{E}_{\text{disc}}[\mathbf{M}] := \sum_{i \in \mathcal{I}} \left(\frac{\bar{Q}_i(\mathbf{M})^2}{M_i^2} + \nu M_i^{2\gamma} \right) L_i, \quad (15)$$

with $\bar{Q}_i(\mathbf{M}) = M_i^2 \frac{(\Delta P)_i}{L_i}$ and the pressure drop $(\Delta P)_i$ again being determined by the Kirchhoff law

$$\sum_{i \in N(j)} (\Delta P)_i \frac{M_i^2}{L_i^2} = S_j \quad \text{for all } j \in \mathcal{V}. \quad (16)$$

The corresponding constrained gradient flow reads now

$$\begin{aligned} \frac{dM_i}{dt} &= 2 \left(\frac{\bar{Q}_i(\mathbf{M})^2}{M_i^3} - \nu \gamma M_i^{2\gamma-1} \right) L_i, \\ &= 2 \left(\frac{(\Delta P)_i^2 M_i}{L_i^2} - \nu \gamma M_i^{2\gamma-1} \right) L_i, \end{aligned} \quad (17)$$

where we again used the fact that $\frac{\partial \bar{Q}_i(\mathbf{M})}{\partial M_i} = 0$. Expressed back in terms of the original variable \mathbf{C} , (17) reads

$$\frac{dC_i}{dt} = 4C_i \left(\frac{Q_i(\mathbf{C})^2}{C_i^2} - \nu \gamma C_i^{\gamma-1} \right) L_i.$$

We observe that the above ODE differs from (12) by the multiplicative factor $4C_i$ in the right-hand side. This is due to the fact that the variable transform changes the geometry of the energy landscape (change of metric in the Onsager structure, see, e.g., [21]). The choice of the “proper” metric would be implied by additional modeling inputs, however, we do not make such a choice here.

Assuming that the network lives in a porous medium, the flux \bar{Q} is given according to Darcy's law (with viscosity scaled to one) by $\bar{Q} = -\mathbb{P}[m]\nabla p = -(m \cdot \nabla p)m$, where p is the continuum pressure. Then, the continuum analogue of the energy (15) is, after rescaling by a multiplicative constant, given by

$$E_0[m] := \frac{1}{2} \int c^2(m \cdot \nabla p[m])^2 + \frac{|m|^{2\gamma}}{\gamma} dx, \quad (18)$$

where we introduced the activation constant $c^2 > 0$, and the continuum analogue of the Kirchhoff's law reads

$$-\nabla \cdot ((m \otimes m)\nabla p[m]) = S. \quad (19)$$

Due to the strong degeneracy of the permeability tensor $m \otimes m$, this equation is in general not solvable. However, let us formally accept its validity for the moment and resolve this issue later. For calculating the gradient flow of (18) with the constraint (19), we consider the first variation in E in the direction φ ,

$$\frac{d}{d\varepsilon} E_0[m + \varepsilon\varphi]_{\varepsilon=0} = \int c^2(m \cdot \nabla p^0)[\varphi \cdot \nabla p^0 + m \cdot \nabla p^1] + |m|^{2(\gamma-1)}m\varphi dx, \quad (20)$$

where we make the expansion $p[m + \varepsilon\varphi] = p^0 + \varepsilon p^1 = p^0[m] + \varepsilon p^1[m, \varphi]$. Inserting this into (19) gives

$$-\nabla \cdot ((m + \varepsilon\varphi) \cdot (\nabla p^0 + \varepsilon\nabla p^1))(m + \varepsilon\varphi) = S.$$

We multiply the above identity by p^0 and integrate by parts. Evaluating the resulting expression at $\varepsilon = 0$, taking into account that $p^0 = p^0[m]$ solves the Poisson equation (19), we obtain

$$\int (m \cdot \nabla p^0)[2\varphi \cdot \nabla p^0 + m \cdot \nabla p^1] dx = 0.$$

Inserting this into (20), we obtain

$$\frac{d}{d\varepsilon} E[m + \varepsilon\varphi]_{\varepsilon=0} = \int -c^2(m \cdot \nabla p^0)(\varphi \cdot \nabla p^0) + |m|^{2(\gamma-1)}m \cdot \varphi dx.$$

Consequently, the formal L^2 -gradient flow of the energy (18) subject to the constraint (19) is given by

$$\partial_t m = c^2(m \cdot \nabla p)\nabla p - |m|^{2(\gamma-1)}m, \quad (21)$$

coupled to the Poisson equation (19). The connection to the discrete setting is as follows: The ODE (17) for $M_i \simeq |m|$ can be seen as a finite difference discretization of (21), and Kirchhoff's law (10) is the corresponding discretization of (19). For

simplicity, let us consider the 1D setting with an equidistant grid $x_j = jL$, $j \in \mathbb{Z}$, where $L > 0$ is the grid size. We identify the set of grid nodes $\{x_j\}_{j \in \mathbb{Z}}$ with the set of vertices \mathcal{V} of the discrete graph. The segments (x_i, x_{i+1}) are identified with the edges $i \in \mathcal{I}$. Due to the invariance of the system (19), (21) with respect to the sign of m , we may without loss of generality assume $m \geq 0$ and identify the discrete values m_i with the scalar M_i . The discrete values of the pressure p_j are defined on the nodes (vertices) x_j . For the discretization of the conductance values m_i , we introduce the dual grid $y_{i+1/2} = \frac{x_i+x_{i+1}}{2}$. Then, (19) is discretized at node x_i as

$$\frac{1}{L} \left(m_{i+1/2}^2 \frac{p_{i+1} - p_i}{L} + m_{i-1/2}^2 \frac{p_{i-1} - p_i}{L} \right) = S_i. \quad (22)$$

With the identification $M_i^2 \simeq m_{i+1/2}^2$, the above formula represents Kirchhoff's law (16) for the 1D equidistant grid. The spatial discretization of (21) at node $y_{i+1/2}$ is given by

$$\frac{dm_{i+1/2}}{dt} = \frac{c^2}{L^2} [m_{i+1/2}(p_{i+1} - p_i)](p_{i+1} - p_i) - |m_{i+1/2}|^{2(\gamma-1)} m_{i+1/2},$$

which can be for $m_{i+1/2} \geq 0$ further rewritten as

$$\frac{dm_{i+1/2}}{dt} = \frac{c^2}{L^2} (p_{i+1} - p_i)^2 m_{i+1/2} - m_{i+1/2}^{2\gamma-1}. \quad (23)$$

This is, up to a possible rescaling of the constants, the ODE (17) for $M_i \simeq m_{i+1/2}$.

As already indicated, contrary to the discrete case (10), the Poisson equation (19) is highly degenerate in directions orthogonal to m and in general unsolvable. To overcome this problem, we introduce an isotropic background permeability of the medium, described by the scalar function $r = r(x) \geq r_0 > 0$, which leads to the modified permeability tensor (6) and inserting into (5), we obtain (7).

2.4 Mesoscopic Models

In order to gain a better understanding of the model structures and the relation between the microscopic and macroscopic scales, we introduce a mesoscopic modeling approach, related to the basic ingredients of the microscopic models. The latter are mainly describing a local equation for the conductivity from a node into certain directions (parametrized by the index $i \in \mathcal{I}$). Let us start with a canonical model of the form

$$\frac{dC_i}{dt} = C_i (A((\Delta P)_i) - R(C_i)),$$

with an activation function A and a relaxation term R . A key observation is to interpret the pressure drop $(\Delta P)_i = Q_i(\mathbf{C})/C_i$ as a difference quotient for a continuous pressure variable p in the direction $\vartheta_i \in \mathcal{S}^1$ of the edge, i.e.

$$(\Delta P)_i \approx L_i \nabla p(x_i) \cdot \vartheta_i,$$

where x_i is the midpoint of the i -th edge. Assuming the existence of a rescaled function

$$B(x_i, \vartheta_i \cdot \nabla p(x_i)) \approx A(L_i \nabla p(x_i) \cdot \vartheta_i),$$

we rewrite the equation for the conductivity as

$$\frac{dC}{dt}(x_i, \vartheta_i) = C(x_i, \vartheta_i)(B(x_i, \vartheta_i \cdot \nabla p(x_i)) - R(C(x_i, \vartheta_i))).$$

This reformulation is the basis for a mesoscopic modeling approach, where we specify the probability density $f(x, \vartheta, C, t)$ to have an edge of conductivity C at x pointing in direction ϑ at time t . Since the conductivity has no sign, we obviously have $f(x, \vartheta, C, t) = f(x, -\vartheta, C, t)$, and it hence suffices to specify f for directions in

$$\mathcal{S}_+^1 = \{\vartheta \in \mathcal{S}^1 \mid \vartheta_1 \geq 0\}.$$

Interpreting the evolution of $C(x, \vartheta)$ as a characteristic equation, we immediately obtain

$$\partial_t f + \partial_C \left(C(B(x, \vartheta \cdot \nabla p(x, t)) - R(C)) f \right) = 0 \quad \text{in } \Omega \times \mathcal{S}_+^1 \times \mathbb{R}_+. \quad (24)$$

The corresponding form of the Poisson equation is given by

$$-\nabla \cdot (\mathbb{P}[f] \nabla p) = S, \quad \mathbb{P}[f](x, t) = \int_{\mathcal{S}_+^1} \int_{\mathbb{R}^+} f(x, \vartheta, C, t) C \vartheta \otimes \vartheta \, dC \, d\vartheta. \quad (25)$$

The system can be understood as a kinetic equation (24) with a nonlinear interaction via the pressure p . We observe that as soon as f is not concentrated in a single direction, the permeability tensor $\mathbb{P}[f]$ is positive definite.

In the case

$$B(x, \vartheta \cdot \nabla p) = c^2 |\vartheta \cdot \nabla p|^2,$$

equation (24) is formally a gradient flow of the form

$$\partial_t f = \partial_C (C f \partial_C \mathcal{F}'[f]) \quad (26)$$

for the energy

$$\mathcal{F}[f] := \frac{1}{2} \int_{\Omega} \int_{\mathbb{R}_+} \int_{\mathcal{S}_+^1} (2\mathfrak{R}(C) + c^2 C |\vartheta \cdot \nabla p[f]|^2) f \, d\vartheta \, dC \, dx,$$

where $\mathfrak{R}'(C) = R(C)$ and $p[f]$ solves (25).

A link to the macroscopic model is provided by defining the conductivity as a moment of the form

$$m(x, t) = \int_{\mathbb{R}_+} \int_{\mathcal{S}_+^1} \sqrt{C} \vartheta f(x, C, \vartheta, t) \, d\vartheta \, dC. \quad (27)$$

Assuming the monokinetic closure

$$f(x, C, \vartheta, t) = \delta(C - \hat{C}(x, t)) \otimes \delta(\vartheta - \hat{\vartheta}(x, t)), \quad (28)$$

for some functions $\hat{C} = \hat{C}(x, t)$, $\hat{\vartheta} = \hat{\vartheta}(x, t)$ and δ the Dirac measure, we obtain $m = \sqrt{\hat{C}} \hat{\vartheta}$. Moreover, a straightforward computation shows that the pressure p satisfies (1) with $r = 0$, and using (24) and integration by parts, we obtain

$$\begin{aligned} \partial_t m &= \partial_t \int_{\mathcal{S}_+^1} \int_{\mathbb{R}_+} \sqrt{C} \vartheta f \, dC \, d\vartheta \\ &= - \int_{\mathcal{S}_+^1} \int_{\mathbb{R}_+} \sqrt{C} \vartheta \partial_C (C(B(x, \vartheta \cdot \nabla p(x, t)) - R(C))f) \, dC \, d\vartheta \\ &= \frac{1}{2} \int_{\mathcal{S}_+^1} \int_{\mathbb{R}_+} \sqrt{C} \vartheta (B(x, \vartheta \cdot \nabla p(x, t)) - R(C))f \, dC \, d\vartheta \\ &= \frac{1}{2} m(B(x, \frac{m}{|m|} \cdot \nabla p) - R(|m|^2)). \end{aligned}$$

With a quadratic choice of B and a power law for R , we obtain a model resembling (1) with $D = 0$. However, a key difference is that the activation direction is proportional to m rather than ∇p , which is due to the fact that the micro- and mesoscopic model never change the direction of an edge, but only increase or decrease its conductivity.

We finally mention that the mesoscopic modeling approach offers several options to naturally incorporate additional effects into the microscopic model, in particular movement of edges and diffusion. This issue is left to future discussion.

3 Mathematical Analysis of the Macroscopic Network Formation PDE System

The main mathematical interest of the PDE system for network formation stems from the highly unusual nonlocal coupling of the elliptic equation (1) for the pressure p to the reaction–diffusion equation (2) for the conductance vector m via the activation term $+c^2(\nabla p \otimes \nabla p)m$ and the latter term’s potential equilibration with the decay term $-|m|^{2(\gamma-1)}m$. A major observation concerning system (1)–(2) is that

it represents the formal $L^2(\Omega)$ -gradient flow associated with the highly nonconvex energy-type functional

$$\mathcal{E}[m] := \frac{1}{2} \int_{\Omega} \left(D^2 |\nabla m|^2 + \frac{|m|^{2\gamma}}{\gamma} + c^2 (m \cdot \nabla p[m])^2 + c^2 r(x) |\nabla p[m]|^2 \right) dx, \quad (29)$$

where $p = p[m] \in H_0^1(\Omega)$ is the unique solution of the Poisson equation (1) with given m , subject to the homogeneous Dirichlet boundary condition on $\partial\Omega$. Note that (29) consists of, respectively, the diffusive energy term, metabolic (relaxation) energy, and the last two terms account for network–fluid interaction energy. We have the energy dissipation:

Lemma 1 (Lemma 1 in [20]) *Let $\mathcal{E}[m^I] < \infty$. Then the energy $\mathcal{E}[m(t)]$ is nonincreasing along smooth solutions of (1)–(2) and satisfies*

$$\frac{d}{dt} \mathcal{E}[m(t)] = - \int_{\Omega} \left(\frac{\partial m}{\partial t}(t, x) \right)^2 dx.$$

As usual, along weak solutions, we obtain the weaker form of energy dissipation,

$$\mathcal{E}[m(t)] \leq \mathcal{E}[m^I],$$

see Theorem 1 below. For convenience of the reader, we reprint the proof of Lemma 1 here.

Proof Multiplication of (1) by p and integration by parts yields

$$\int_{\Omega} (r|\nabla p|^2 + |m \cdot \nabla p|^2 - pS) dx = 0.$$

Subtracting the c^2 -multiple of the above identity from (29), we obtain

$$\mathcal{E}[m(t)] = \frac{1}{2} \int_{\Omega} \left(D^2 |\nabla m|^2 + \frac{|m|^{2\gamma}}{\gamma} - c^2 |m \cdot \nabla p|^2 - c^2 r |\nabla p|^2 + 2c^2 pS \right) dx,$$

so that, after integration by parts in suitable terms,

$$\begin{aligned} \frac{d\mathcal{E}}{dt}[m(t)] &= - \int_{\Omega} D^2 \Delta m \cdot \partial_t m dx + \int_{\Omega} |m|^{2(\gamma-1)} m \cdot \partial_t m dx - c^2 \int_{\Omega} (m \cdot \nabla p) \nabla p \cdot \partial_t m dx \\ &\quad + c^2 \int_{\Omega} \nabla \cdot [(m \cdot \nabla p)m] \partial_t p dx + c^2 \int_{\Omega} r(\Delta p)(\partial_t p) dx + c^2 \int_{\Omega} (\partial_t p) S dx \\ &= - \int_{\Omega} \left[D^2 \Delta m - |m|^{2(\gamma-1)} m + c^2 (m \cdot \nabla p) \nabla p \right] \cdot \partial_t m dx \end{aligned}$$

$$\begin{aligned}
& +c^2 \int_{\Omega} [\nabla \cdot (r \nabla p + (m \otimes m) \nabla p) + S] \partial_t p \, dx \\
& = - \int_{\Omega} |\partial_t m|^2 \, dx.
\end{aligned}$$

□

In the next section we give an overview of the existence proof for global weak solutions of the system (1)–(2) for $\gamma \geq 1/2$. As we point out in Remark 3, the borderline case $\gamma = 1/2$ requires a slightly different treatment, since the algebraic term in (1) formally becomes $m/|m|$, and an interpretation has to be given for $m = 0$. Mild solutions with $m(t) \in (L^\infty(\Omega))^d$ can be built by a perturbation method, see Section 3.2, but, in general, only locally in time. Then, in Section 3.3, we construct stationary solutions of the system (1)–(2) with the simplifying assumption $D = 0$ and study their stability properties. Special attention is paid to the case $\gamma = 1$ since it leads to an interesting, highly nonlinear Poisson-type equation.

For simplicity and without loss of generality, we will consider constant background permeability $r = r(x) \equiv 1$ in the rest of Section 3. Moreover, we will adopt the usual convention that generic, not necessarily equal, constants will be denoted by C . We will only make specific use of the Poincaré constant C_Ω , i.e.,

$$\|u\|_{L^2(\Omega)} \leq C_\Omega \|\nabla u\|_{L^2(\Omega)} \quad \text{for all } u \in H_0^1(\Omega). \quad (30)$$

3.1 Global Existence of Weak Solutions for $\gamma \geq 1/2$

Our first goal is to prove the existence of global weak solutions of the system (1)–(4). The major mathematical difficulty is that a priori estimates for m are too weak to use elliptic regularizing effects for p . Therefore, weak solutions belong just to the energy space that we define below. We first prove the result for $\gamma > 1/2$ and comment on the borderline case $\gamma = 1/2$ in Remark 3.

Theorem 1 (Weak solutions) *Let $\gamma > 1/2$, $S \in L^2(\Omega)$ and $m^I \in H_0^1(\Omega)^d \cap L^{2\gamma}(\Omega)^d$. Then the problem (1)–(4) admits a global weak solution $(m, p[m])$ with $\mathcal{E}[m] \in L^\infty(0, \infty)$, i.e., with*

$$\begin{aligned}
m & \in L^\infty(0, \infty; H_0^1(\Omega)) \cap L^\infty(0, \infty; L^{2\gamma}(\Omega)), \quad \partial_t m \in L^2((0, \infty) \times \Omega), \\
\nabla p & \in L^\infty(0, \infty; L^2(\Omega)), \quad m \cdot \nabla p \in L^\infty(0, \infty; L^2(\Omega)).
\end{aligned}$$

These solutions satisfy the energy inequality, with \mathcal{E} given by (29),

$$\mathcal{E}[m(t)] + \int_0^t \int_{\Omega} \left(\frac{\partial m}{\partial t}(s, x) \right)^2 \, dx \, ds \leq \mathcal{E}[m^I] \quad \text{for all } t \geq 0. \quad (31)$$

We proceed by proving existence of solutions of a regularized problem and a subsequent limit passage. For this, we need several auxiliary analytical results that we present below; for their proofs, we refer to [20].

We first consider the semilinear parabolic problem on Ω

$$\frac{\partial m}{\partial t} - D^2 \Delta m + |m|^{2(\gamma-1)} m = f \quad (32)$$

subject to the initial and boundary conditions

$$m(t=0) = m^I \quad \text{in } \Omega, \quad m = 0 \quad \text{on } \partial\Omega. \quad (33)$$

Lemma 2 *For every $D > 0$, $\gamma > 1/2$, $T > 0$ and $f \in L^2((0, T) \times \Omega)^d$, the problem (32)–(33) with $m^I \in H_0^1(\Omega)^d$ admits a unique weak solution $m \in L^\infty(0, T; H_0^1(\Omega)^d \cap L^2(0, T; H^2(\Omega))^d \cap L^\infty(0, T; L^{2\gamma}(\Omega))^d$ with $\partial_t m \in L^2((0, T) \times \Omega)^d$ and*

$$\|m\|_{L^\infty(0, T; H_0^1(\Omega))} \leq C \left(\|f\|_{L^2((0, T) \times \Omega)} + \|m^I\|_{H_0^1(\Omega)} \right), \quad (34)$$

$$\|\Delta m\|_{L^2((0, T) \times \Omega)} \leq C \left(\|f\|_{L^2((0, T) \times \Omega)} + \|m^I\|_{H_0^1(\Omega)} \right). \quad (35)$$

We will also need the following Lemma concerning the algebraic term $|m|^{2(\gamma-1)} m$.

Lemma 3 *Fix $\gamma > 1/2$ and let the sequence $\{m^k\}_{k \in \mathbb{N}}$ be uniformly bounded in $L^{2\gamma}((0, T) \times \Omega)$ and converging to m in the norm topology of $L^2((0, T) \times \Omega)$ as $k \rightarrow \infty$. Then, for every test function $\varphi \in C_c^\infty([0, T) \times \Omega)$,*

$$\int_0^T \int_\Omega |m^k|^{2(\gamma-1)} m^k \varphi \, dx \, dt \rightarrow \int_0^T \int_\Omega |m|^{2(\gamma-1)} m \varphi \, dx \, dt \quad \text{as } k \rightarrow \infty.$$

Next, we study the properties of solutions of the regularized Poisson equation

$$-\nabla \cdot [\nabla p + \bar{m}(\bar{m} \cdot \nabla p) * \eta] = S, \quad (36)$$

where \bar{m} and S are given functions on Ω and $\eta = \eta(|x|) \in C^\infty(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ is a smooth mollifier with real, nonnegative Fourier transform (in particular, we shall use the heat kernel later on). The convolution $f * \eta$ for $f \in L^1(\Omega)$ is defined as

$$f * \eta(x) = \int_{\mathbb{R}^d} f(y) \eta(x-y) \, dy,$$

where we extend f by zero to \mathbb{R}^d , i.e., $f(y) = 0$ for $y \in \mathbb{R}^d \setminus \Omega$. We need the following technical Lemma that follows from a simple application of the Parseval identity.

Lemma 4 For any $u \in L^1(\mathbb{R}^d)$ and $\eta \in L^\infty(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ with nonnegative Fourier transform $\hat{\eta} \geq 0$ on \mathbb{R}^d , the identity holds

$$\int_{\mathbb{R}^d} (u * \eta) u \, dx = \int_{\mathbb{R}^d} |u * \rho|^2 \, dx \geq 0,$$

where ρ is the inverse Fourier transform of $(\hat{\eta})^{1/2}$.

Next, we have the following two results concerning the existence and uniqueness of solutions of the regularized Poisson equation and their stability.

Lemma 5 For every $\bar{m} \in L^2(\Omega)$ and $S \in L^2(\Omega)$, the regularized Poisson equation (36) has a unique weak solution $p \in H_0^1(\Omega)$. Moreover,

$$\|\nabla p\|_{L^2(\Omega)} \leq C_\Omega \|S\|_{L^2(\Omega)}, \quad (37)$$

where C_Ω is the Poincaré constant defined in (30).

Lemma 6 Let \bar{m}^k be a sequence of functions converging to \bar{m} in the norm topology of $L^2(\Omega)$, and denote by $p^k \in H_0^1(\Omega)$ the corresponding weak solutions of (36) subject to homogeneous Dirichlet boundary conditions. Then, p^k converges strongly in $H^1(\Omega)$ to the unique solution p of (36) as $k \rightarrow \infty$.

Now we are ready to consider, for $\varepsilon > 0$, the perturbed problem

$$-\nabla \cdot [\nabla p + m(m \cdot \nabla p) * \eta_\varepsilon] = S, \quad (38)$$

$$\frac{\partial m}{\partial t} - D^2 \Delta m - c^2 [(m \cdot \nabla p) * \eta_\varepsilon] \nabla p + |m|^{2(\gamma-1)} m = 0, \quad (39)$$

with $(\eta_\varepsilon)_{\varepsilon>0}$ the d -dimensional heat kernel $\eta_\varepsilon(x) = (4\pi\varepsilon)^{-d/2} \exp(-|x|^2/4\varepsilon)$, on a bounded domain $\Omega \subset \mathbb{R}^d$, $d \leq 3$, with smooth boundary $\partial\Omega$, subject to homogeneous Dirichlet boundary conditions on $\partial\Omega$ for m and p ,

$$m(t, x) = 0, \quad p(t, x) = 0 \quad \text{for all } x \in \partial\Omega, \quad t \geq 0, \quad (40)$$

and the initial condition for m ,

$$m(t=0, x) = m^I(x) \quad \text{for all } x \in \Omega. \quad (41)$$

Let us note that the heat kernel η_ε satisfies the assumptions of Lemma 4 for any $\varepsilon > 0$. Since the following result regarding the existence of weak solutions of the perturbed problem is central for the analysis in this section, we provide its proof for convenience of the reader.

Theorem 2 (Existence for the perturbed model) Let $m^I \in L^2(\Omega)^d$. For any $\varepsilon > 0$ there exists a weak solution (m, p) of the system (38)–(41) with $p \in L^\infty(0, T; H_0^1(\Omega))$ and $m \in L^2(0, T; H_0^1(\Omega))^d$, $\Delta m \in L^2((0, T) \times \Omega)^d$ and $\partial_t m \in L^2((0, T) \times \Omega)^d$.

Proof We shall employ the Leray–Schauder fixed point theorem. We fix $\varepsilon > 0$ and construct the mapping $\Phi : \bar{m} \mapsto m$ in two steps: For a given $\bar{m} \in L^2((0, T) \times \Omega)$, we set p to be the unique weak solution of

$$-\nabla \cdot [\nabla p + \bar{m}(\bar{m} \cdot \nabla p) * \eta_\varepsilon] = S, \quad (42)$$

constructed in Lemma 5 (to be precise, we use a straightforward modification of Lemma 5 where the Lax–Milgram theorem is applied in the space $L^2(0, T; H_0^1(\Omega))$). By the same Lemma, we have the a priori estimate $\|\nabla p\|_{L^\infty(0, T; L^2(\Omega))} \leq C$ independent of \bar{m} . We set $q_\varepsilon := (\nabla p \cdot \bar{m}) * \eta_\varepsilon$ and note that it is a priori bounded in $L^2(0, T; L^\infty(\Omega))$ due to the Young inequality

$$\begin{aligned} \|q_\varepsilon\|_{L^2(0, T; L^\infty(\Omega))} &\leq \|\eta_\varepsilon\|_{L^\infty(\mathbb{R}^d)} \|\nabla p \cdot \bar{m}\|_{L^2(0, T; L^1(\Omega))} \\ &\leq C_\varepsilon \|\nabla p\|_{L^\infty(0, T; L^2(\Omega))} \|\bar{m}\|_{L^2((0, T) \times \Omega)}. \end{aligned} \quad (43)$$

Then, we employ Lemma 2 with $f := c^2 q_\varepsilon \nabla p \in L^2((0, T) \times \Omega)$ and set $\Phi(\bar{m}) := m \in L^2(0, T; H_0^1(\Omega))^d$ to be the unique weak solution of

$$\frac{\partial m}{\partial t} - D^2 \Delta m + |m|^{2(\gamma-1)} m = c^2 q_\varepsilon \nabla p \quad (44)$$

subject to the initial condition $m(t = 0) = m^I$. The Lemma provides the estimate

$$\begin{aligned} \|m\|_{L^\infty(0, T; L^2(\Omega))} + \|\nabla m\|_{L^2((0, T) \times \Omega)} &\leq C \left(\|m^I\|_{L^2(\Omega)} + c^2 \|q_\varepsilon \nabla p\|_{L^2((0, T) \times \Omega)} \right) \\ &\leq C \left(\|m^I\|_{L^2(\Omega)} + c^2 \|q_\varepsilon\|_{L^2(0, T; L^\infty(\Omega))} \|\nabla p\|_{L^\infty(0, T; L^2(\Omega))} \right) \\ &\leq \tilde{C} + C_\varepsilon \|\nabla p\|_{L^\infty(0, T; L^2(\Omega))}^2 \|\bar{m}\|_{L^2((0, T) \times \Omega)} \end{aligned}$$

for suitable constants $\tilde{C}, C_\varepsilon > 0$. This also implies an a priori bound on $\partial_t m$ in $L^2(0, T; H^{-1}(\Omega))$, so that $\Phi : \bar{m} \mapsto m$ maps bounded sets in $L^2((0, T) \times \Omega)$ onto relatively compact ones.

To prove continuity of Φ , consider a sequence \bar{m}^k converging to \bar{m} in the norm topology of $L^2((0, T) \times \Omega)$. Due to Lemma 5, the sequence ∇p^k of the corresponding solutions of (42) converges weakly-* in $L^\infty(0, T; L^2(\Omega))$ to some ∇p . The bound (43) allows to extract a subsequence of $q_\varepsilon^k := (\bar{m}^k \cdot \nabla p^k) * \eta_\varepsilon$ converging weakly in $L^2(0, T; L^\infty(\Omega))$ to $q_\varepsilon := (\bar{m} \cdot \nabla p) * \eta_\varepsilon$. Therefore, we can pass to the limit in the weak formulation of the regularized Poisson equation (42) and conclude that p is its unique solution corresponding to \bar{m} . To pass to the limit in (44), we use strong convergence of ∇p^k in $L^2((0, T) \times \Omega)$ provided by Lemma 6 (strictly speaking, by its straightforward modification for time-dependent functions \bar{m}^k). Then, the limit passage in the term $q_\varepsilon^k \nabla p^k$ is straightforward. For the algebraic term $|m^k|^{2(\gamma-1)} m^k$, we employ Lemma 3 and finally conclude, by the uniqueness of solutions of (44), the continuity of the mapping $m = \Phi(\bar{m})$.

Finally, we have to prove that the set $\mathcal{Y} := \{m \in L^2((0, T) \times \Omega); \kappa\Phi(m) = m \text{ for some } 0 < \kappa \leq 1\}$ is bounded. Note that the elements of this set are weak solutions of the system

$$\begin{aligned} -\nabla \cdot [\nabla p + m(m \cdot \nabla p) * \eta_\varepsilon] &= S, \\ \frac{\partial m}{\partial t} - D^2 \Delta m - \kappa c^2 [(m \cdot \nabla p) * \eta_\varepsilon] \nabla p + \kappa^{-2(\gamma-1)} |m|^{2(\gamma-1)} m &= 0, \end{aligned}$$

subject to homogeneous Dirichlet boundary conditions for m and p on $\partial\Omega$ and the initial condition $m(t = 0, x) = \kappa m^I(x)$ in Ω . Multiplication of the first equation by $c^2 \kappa p$ and of the second equation by m , integration by parts and subtraction of the two identities yields

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} |m|^2 dx + D^2 \int_{\Omega} |\nabla m|^2 dx + \kappa^{-2(\gamma-1)} \int_{\Omega} |m|^{2\gamma} dx \\ = c^2 \kappa \left(\int_{\Omega} p S dx - \int_{\Omega} |\nabla p|^2 dx \right). \end{aligned}$$

This implies, for any $0 < \kappa \leq 1$,

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} |m|^2 dx + D^2 \int_{\Omega} |\nabla m|^2 dx &\leq c^2 \kappa \left(\|S\|_{L^2(\Omega)} \|p\|_{L^2(\Omega)} - \|\nabla p\|_{L^2(\Omega)}^2 \right) \\ &\leq c^2 \kappa C_{\Omega}^2 \|S\|_{L^2(\Omega)}^2, \end{aligned}$$

where we used the Poincaré inequality (30) with the constant C_{Ω} . This immediately gives the a priori boundedness of m in $L^\infty(0, T; L^2(\Omega))$ and thus the boundedness of the set \mathcal{Y} .

Finally, an application of Lemma 2 with $f := q_\varepsilon \nabla p \in L^2((0, T) \times \Omega)^d$ implies that $\Delta m \in L^2((0, T) \times \Omega)^d$ and $\partial_t m \in L^2((0, T) \times \Omega)^d$. \square

Finally, we shall pass to the limit $\varepsilon \rightarrow 0$ in (38)–(41) and obtain a global solution of the system (1)–(4). The main tool is the dissipation of the modified energy

$$\mathcal{E}_\varepsilon[m] := \frac{1}{2} \int \left(D^2 |\nabla m|^2 + \frac{|m|^{2\gamma}}{\gamma} + c^2 (m \cdot \nabla p) [(m \cdot \nabla p) * \eta_\varepsilon] + c^2 |\nabla p|^2 \right) dx. \quad (45)$$

Note that by Lemma 4, we have

$$\int_{\Omega} (m \cdot \nabla p) [(m \cdot \nabla p) * \eta_\varepsilon] dx = \int_{\mathbb{R}^d} |(m \cdot \nabla p) * \rho_\varepsilon|^2 dx \geq 0$$

with $\hat{\eta}_\varepsilon = |\hat{\rho}_\varepsilon|^2$, so that $\mathcal{E}_\varepsilon[m(t)] \geq 0$.

Lemma 7 *Let (m, p) be a solution of (38)–(39) constructed in Theorem 2 and assume $\mathcal{E}_\varepsilon[m^I] < \infty$. Then, the energy (45) satisfies*

$$\mathcal{E}_\varepsilon[m(t)] + \int_0^t \int_\Omega \left(\frac{\partial m}{\partial t}(s, x) \right)^2 dx ds = \mathcal{E}_\varepsilon[m^I] \quad \text{for all } 0 \leq t \leq T. \quad (46)$$

The proof is an adaptation of the formal proof of Lemma 1, see [20]. We are now ready to pass to the limit $\varepsilon \rightarrow 0$ in (38)–(39). Again, since the limit passage is an essential step in the analysis of this section, we provide the full proof for convenience of the reader.

Lemma 8 *Let $(m^\varepsilon, p^\varepsilon)_{\varepsilon > 0}$ be a family of weak solution of (38)–(39) constructed in Theorem 2 and assume $\mathcal{E}[m^I] < \infty$. Then, there exists a subsequence converging to (m, p) as $\varepsilon \rightarrow 0$, where (m, p) is a weak solution of (1)–(4), satisfying the energy dissipation inequality (31).*

Proof Note that the Poisson equation (38) at $t = 0$ implies

$$\int_\Omega |\nabla p^\varepsilon[m^I]|^2 dx + \int_\Omega |(m^I \cdot \nabla p^\varepsilon[m^I]) * \rho_\varepsilon|^2 dx = \int_\Omega p^\varepsilon[m^I] S dx, \quad (47)$$

such that by Lemma 5 $\mathcal{E}_\varepsilon[m^I]$ is uniformly bounded as $\varepsilon \rightarrow 0$. Then, the energy dissipation given by Lemma 7 provides the following uniform a priori estimates,

$$\begin{aligned} m^\varepsilon &\in L^\infty(0, T; H_0^1(\Omega)) \cap L^\infty(0, T; L^{2\gamma}(\Omega)), \quad \partial_t m^\varepsilon \in L^2((0, T) \times \Omega), \\ \nabla p^\varepsilon &\in L^\infty(0, T; L^2(\Omega)), \quad (m^\varepsilon \cdot \nabla p^\varepsilon) * \rho_\varepsilon \in L^\infty(0, T; L^2(\mathbb{R}^d)), \end{aligned}$$

with $\hat{\eta}_\varepsilon = |\hat{\rho}_\varepsilon|^2$. The last bound implies a uniform estimate on $q_\varepsilon := (m^\varepsilon \cdot \nabla p^\varepsilon) * \eta_\varepsilon$ in $L^\infty(0, T; L^2(\Omega))$. Indeed, taking any test function $\varphi \in L^1(0, T; L^2(\Omega))$, we have

$$\begin{aligned} \int_0^T \int_\Omega q_\varepsilon \varphi dx dt &= \int_0^T \int_{\mathbb{R}^d} [(m^\varepsilon \cdot \nabla p^\varepsilon) * \rho_\varepsilon][\varphi * \rho_\varepsilon] dx dt \\ &\leq \int_0^T \| (m^\varepsilon \cdot \nabla p^\varepsilon) * \rho_\varepsilon \|_{L^2(\mathbb{R}^d)} \| \varphi * \rho_\varepsilon \|_{L^2(\mathbb{R}^d)} dt \\ &\leq \| (m^\varepsilon \cdot \nabla p^\varepsilon) * \rho_\varepsilon \|_{L^\infty(0, T; L^2(\mathbb{R}^d))} \int_0^T \| \varphi \|_{L^2(\Omega)} \| \rho_\varepsilon \|_{L^1(\mathbb{R}^d)} dt \\ &= \| (m^\varepsilon \cdot \nabla p^\varepsilon) * \rho_\varepsilon \|_{L^\infty(0, T; L^2(\mathbb{R}^d))} \| \varphi \|_{L^1(0, T; L^2(\Omega))}, \end{aligned}$$

where we used the fact that, by definition of ρ_ε , $\| \rho_\varepsilon \|_{L^1(\mathbb{R}^d)} = 1$ for every $\varepsilon > 0$. Therefore, by duality, q_ε is uniformly bounded in $L^\infty(0, T; L^2(\Omega))$ and there exists a subsequence converging weakly-* in this space to some $q \in L^\infty(0, T; L^2(\Omega))$. We note that due to the compact embedding (Corollary 4 in [43]), a subsequence of m^ε converges to some m in the norm topology of $L^2((0, T) \times \Omega)$. Then, a slight modification of Lemma 6 provides the strong convergence of p^ε to p in $L^2(0, T; H_0^1(\Omega))$, where p is the unique solution of the Poisson equation (1) with m . Consequently, the product $m^\varepsilon \cdot \nabla p^\varepsilon$ converges strongly to $m \cdot \nabla p$ in $L^1((0, T) \times \Omega)$, and for every test function $\varphi \in C_0^\infty([0, T) \times \Omega)$ we have

$$\int_0^T \int_{\mathbb{R}^d} q_\varepsilon \varphi \, dx \, dt = \int_0^T \int_{\mathbb{R}^d} (m^\varepsilon \cdot \nabla p^\varepsilon)(\varphi * \eta_\varepsilon) \, dx \xrightarrow{\varepsilon \rightarrow 0} \int_0^T \int_{\mathbb{R}^d} (m \cdot \nabla p) \varphi \, dx,$$

where we used the fact that $\varphi * \eta_\varepsilon$ converges to φ in $C([0, T] \times \bar{\Omega})$ as $\varepsilon \rightarrow 0$ due to the Arzelà–Ascoli theorem. Therefore, we identify the limit $q = m \cdot \nabla p$.

We are now ready to pass to the limit in the weak formulation of the nonlinear terms of (38)–(39). The term $q_\varepsilon m^\varepsilon$ in (38) converges to $(m \cdot \nabla p)m$ due to the weak-* convergence of q_ε in $L^\infty(0, T; L^2(\Omega))$ and strong convergence of m^ε in $L^2((0, T) \times \Omega)$. The term $q_\varepsilon \nabla p^\varepsilon$ in (39) converges to $(m \cdot \nabla p)\nabla p$ due to the strong convergence of ∇p^ε in $L^2((0, T) \times \Omega)$. Finally, the limit passage in the term $|m^\varepsilon|^{2(\gamma-1)}m^\varepsilon$ is provided by Lemma 3 due to the uniform boundedness of m^ε in $L^{2\gamma}((0, T) \times \Omega)$.

The energy dissipation inequality (31) follows from (46) due to the weak lower semicontinuity of terms defining $\mathcal{E}[m]$ and from the fact that $\mathcal{E}_\varepsilon[m^\varepsilon] \rightarrow \mathcal{E}[m^I]$ as $\varepsilon \rightarrow 0$. Indeed, Lemma (6) provides strong converge of $\nabla p^\varepsilon[m^I]$ to $\nabla p[m^I]$ in $L^2(\Omega)^d$. Due to the embedding of $H_0^1(\Omega)$ into $L^6(\Omega)$ for $d \leq 3$, the term $m^I \cdot \nabla p^\varepsilon[m^I]$ converges strongly in $L^{3/2}(\Omega)$ to $m^I \cdot \nabla p[m^I]$. Consequently, the limit passage $\varepsilon \rightarrow 0$ in the identity (47) gives $\mathcal{E}_\varepsilon[m^I] \rightarrow \mathcal{E}[m^I]$. \square

To conclude the proof of Theorem 1, we fix a $T > 0$ and construct a global solution $(m, p[m])$ on $(0, \infty)$ by concatenation of weak solutions on time intervals of length T as constructed in Lemma 8. This is possible due to the energy dissipation inequality (31) and yields the global solution announced in Theorem 1.

Remark 2 Since the solution (m, p) constructed in Theorem 1 satisfies $m \cdot \nabla p \in L^\infty(0, \infty; L^2(\Omega))$ and $\nabla p \in L^\infty(0, \infty; L^2(\Omega))$, implying $(m \cdot \nabla p)\nabla p \in L^\infty(0, \infty; L^1(\Omega))$ and $\partial_t m \in L^2((0, \infty) \times \Omega)$, $|m|^{2\gamma-1} \in L^\infty(0, \infty; L^{2\gamma/(2\gamma-1)})$, we conclude $\Delta m \in L^2(0, \infty; L^1(\Omega))$. Theorems 1.7 and 3.3 in [19] imply Besov regularity for m , namely $m \in L^2(0, \infty; B_\infty^{2,1}(\Omega))$. Also, weak solutions satisfy the equation (1) pointwise almost everywhere (while no regularity on second derivatives of p is guaranteed). Finally, we note that—by the same line of argument—weak stationary solutions possess the Besov regularity $m \in B_\infty^{2,1}(\Omega)$.

Remark 3 The case $\gamma = 1/2$ requires special care since the algebraic term in (1) formally becomes $m/|m|$ and an interpretation has to be given for $m = 0$. In particular, (1) has to be substituted by the differential inclusion

$$\partial_t m - D^2 \Delta m - c^2(m \cdot \nabla p[m])\nabla p[m] \in -\partial \mathcal{R}(m), \quad (48)$$

where $\partial \mathcal{R}$ is the subdifferential of $\mathcal{R}(m) := \int_\Omega |m| \, dx$, in particular,

$$\begin{aligned} \partial \mathcal{R}(m) = \{u \in L^\infty(\Omega)^d; u(x) &= m(x)/|m(x)| \text{ if } m(x) \neq 0, \\ |u(x)| &\leq 1 \text{ if } m(x) = 0\}. \end{aligned}$$

Then, the statement of Theorem 1 remains valid for the system (1), (48), see [22] for details. We conjecture that m is in fact a slow solution of (48), i.e., that it solves

$$\partial_t m - D^2 \Delta m - c^2(m \cdot \nabla p[m]) \nabla p[m] = r(m)$$

with $r(m)$ given by

$$[r(m)](x) = \begin{cases} m(x)/|m(x)| & \text{when } m(x) \neq 0, \\ 0 & \text{when } m(x) = 0. \end{cases}$$

A proof of this conjecture remains an open problem.

3.2 Local Existence and Uniqueness of Mild Solutions for $\gamma > 1/2$

In this section, we provide the main result concerning local in time existence and uniqueness of mild solutions of the system (1)–(4). We will work with the slightly unusual space of functions with vanishing mean oscillation on Ω , $\text{VMO}(\Omega)$. We provide its definition for the convenience of the reader and refer to [8, 41] for more details.

A locally integrable function f on Ω belongs to the John-Nirenberg space $\text{BMO}(\Omega)$ of functions of bounded mean oscillation [27] if

$$\sup_B \frac{1}{|B|} \int_B |f(x) - f_B| dx < +\infty,$$

where B ranges in the class of the balls contained in Ω and $f_B = \frac{1}{|B|} \int_B f(x) dx$. A very important subspace of BMO , introduced by L. Sarason [41], consists of functions of vanishing mean oscillation, VMO . If $f \in \text{BMO}(\Omega)$, set

$$\zeta(r) := \sup_{s \leq r} \frac{1}{|B_s|} \int_{B_s} |f(x) - f_{B_s}| dx,$$

where this time B_s are balls of radius s . We say that $f \in \text{VMO}(\Omega)$ if, in addition, $\lim_{r \rightarrow 0} \zeta(r) = 0$.

We fix $T > 0$ and define the Banach spaces $\mathbb{X} := (L^\infty(\Omega) \cap \text{VMO}(\Omega))^d$, equipped with the L^∞ -norm, and $\mathcal{X}_T := L^\infty(0, T; \mathbb{X})$. The reason why we work in the less usual space \mathbb{X} is the following elliptic regularity result of [13].

Lemma 9 *Let $m \in \mathbb{X}$, $f \in L^q(\Omega)$ for some $1 \leq q < \infty$ and $S \in L^r(\Omega)$ with $r = \max\{1, dq/(d+q)\}$. Then, the PDE*

$$\begin{aligned} -\nabla \cdot ((I + m \otimes m) \nabla p + f) &= S && \text{in } \Omega, \\ p &= 0 && \text{on } \partial\Omega \end{aligned}$$

has a unique weak solution $p \in W^{1,q}(\Omega)$ and there exists a constant $C(\|m\|_{\mathbb{X}}) > 0$, independent of f and S , such that

$$\|\nabla p\|_{L^q(\Omega)} \leq C(\|m\|_{\mathbb{X}}) (\|f\|_{L^q(\Omega)} + \|S\|_{L^r(\Omega)}). \quad (49)$$

We denote $L := D^2\Delta$, where Δ stands for the Dirichlet Laplacian on Ω . Moreover, we define the mapping \mathcal{T} on $\mathbb{R} \times \mathcal{X}_T$ by

$$\mathcal{T}(c^2, m) = e^{Lt} m^I + \int_0^t e^{L(t-s)} (c^2 F[m](s) - G[m](s)) ds \quad (50)$$

with $F[m] = (m \cdot \nabla p[m]) \nabla p[m]$ where $p[m]$ is the $H_0^1(\Omega)$ -solution of the Poisson equation (1) with m given, and $G[m] = |m|^{2\gamma-1}m$.

Obviously, (m, p) is a mild solution of the system (1)–(4) with the activation parameter $c^2 \geq 0$ subject to the initial datum m^I if m is a fixed point of \mathcal{T} , i.e., $\mathcal{T}(c^2, m) = m$. For $c^2 = 0$, we have the trivial fixed point $(0, m_0 = e^{Lt} m^I)$. Our main result below provides the existence of an unbounded continuum of nontrivial mild solutions, i.e., fixed points of \mathcal{T} with $c^2 > 0$. We refer to [22] for its proof.

Theorem 3 *Let $\gamma > 1/2$, $m^I \in \mathbb{X}$ and $S \in L^\infty(\Omega)$. Then, there exists an unbounded continuum of solutions (λ, m) of $\mathcal{T}(\lambda, m) = m$ in $[0, \infty) \times \mathcal{X}_T$ emanating from $(0, m_0 = e^{Lt} m^I)$. Moreover, if $\gamma \geq 1$, the mild solutions are unique locally in time.*

Remark 4 The assertion of Theorem 3 implies the following: If for some $c^2 > 0$, there is no fixed point of \mathcal{T} in \mathcal{X}_T , then there exists a *bounded* sequence of $c_k^2 > 0$ and a sequence of corresponding fixed points $m^k \in \mathcal{X}_T$ of $\mathcal{T}(c_k^2, \cdot)$, such that $\|m^k\|_{\mathcal{X}_T} \rightarrow \infty$ as $k \rightarrow \infty$. Moreover, for $\gamma \geq 1$, the fixed points m of $\mathcal{T}(c^2, m)$ are either global in time classical solutions of (1)–(4), or there exists a $T > 0$ and a sequence $t_k \rightarrow T$ as $k \rightarrow \infty$ such that $\|m\|_{C([0, t_k]; \mathbb{X})} \rightarrow \infty$ as $k \rightarrow \infty$.

Remark 5 In the two-dimensional setting $d = 2$, it is possible to apply Theorem 3 in the space $L^\infty((0, T) \times \Omega)^d$ instead of \mathcal{X}_T . The estimate of Lemma 9 is replaced by the Meyers estimate, Theorem 1 in [32], which states that (49) holds for *some* $q > 2$ if m is bounded in $L^\infty((0, T) \times \Omega)^d$.

Remark 6 In the one-dimensional setting $d = 1$, the branch of solutions constructed in Theorem 3 is in fact global in $c^2 \geq 0$ for every $T > 0$. This follows from the L^∞ bound on $\partial_x p$ provided by Lemma 10 below. Then, the maximum principle yields an a priori bound on m in \mathcal{X}_T for every $T > 0$. In other words, a unique global in time mild solution exists for every value $c^2 \geq 0$ and every $m^I \in L^\infty(0, 1)$.

Lemma 10 *Let $f \in L^1(0, 1)$ and b measurable on $(0, 1)$ such that $b(x) \geq b_0 > 0$ for all $x \in [0, 1]$. Let $p \in H_0^1(0, 1)$ be the unique weak solution of*

$$-\partial_x(b(x)\partial_x p(x)) = f(x) \quad (51)$$

on $[0, 1]$ subject to the boundary conditions $p(0) = p(1) = 0$. Then, the estimate holds

$$|\partial_x p(x)| \leq \frac{2 \|f\|_{L^1(0,1)}}{b(x)} \quad \text{for all } x \in (0, 1).$$

Proof We assume b smooth enough and integrate (51) on $(0, x)$,

$$b(x)\partial_x p(x) = -F(x) + B,$$

where $F(x) = \int_0^x f(s) ds$ and B is an integration constant. Dividing by $b(x)$ and integrating once again leads to

$$p(x) = - \int_0^x \frac{F(s)}{b(s)} ds + B \int_0^x \frac{ds}{b(s)}.$$

The right boundary condition $p(1) = 0$ gives the value for B ,

$$B = \left(\int_0^1 \frac{F(s)}{b(s)} ds \right) \left(\int_0^1 \frac{ds}{b(s)} \right)^{-1},$$

which immediately shows $|B| \leq \|F\|_{L^\infty(0,1)}$. Using this in the above formula for $\partial_x p$ yields

$$|\partial_x p(x)| \leq \frac{|F(x)|}{b(x)} + \frac{|B|}{b(x)} \leq \frac{2 \|F\|_{L^\infty(0,1)}}{b(x)} \leq \frac{2 \|f\|_{L^1(0,1)}}{b(x)}$$

and a density argument finishes the proof. \square

3.3 Analysis of Steady States for $D = 0$

In this section, we calculate steady states of the PDE system (1)–(2) under the simplifying assumption $D = 0$, which allows to obtain explicit formulae. We first consider the spatially one-dimensional setting, where we are able to analyze the nonlinear stability of the steady states. In the multidimensional setting, we are merely able to construct *pointwise* stationary solutions of (1)–(2). Finally, we focus on the case $\gamma = 1$, which leads to an interesting, highly nonlinear Poisson-type equation.

3.3.1 Nonlinear Stability Analysis in 1d

We consider the system (1)–(2) in the spatially one-dimensional setting and without loss of generality we set $\Omega := (0, 1)$. Then, the system with $D = 0$ reads

$$-\partial_x(\partial_x p + m^2 \partial_x p) = S, \quad (52)$$

$$\partial_t m - c^2 (\partial_x p)^2 m + |m|^{2(\gamma-1)} m = 0, \quad (53)$$

Additionally, throughout this section, we assume $S > 0$ a.e. on $(0, 1)$, and for mathematical convenience, we prescribe the mixed boundary conditions for p ,

$$\partial_x p(0) = 0, \quad p(1) = 0,$$

and homogeneous Neumann boundary condition for m . Integrating (52) with respect to x , we obtain

$$(1 + m^2) \partial_x p = - \int_0^x S(y) dy.$$

Denoting $B(x) := \int_0^x S(y) dy > 0$ for $x \in (0, 1)$, we have

$$\partial_x p = - \frac{B(x)}{1 + m^2}, \quad (54)$$

so that the system (52)–(53) is rewritten as the family of ODEs

$$\partial_t m = \left(\frac{c^2 B(x)^2}{(1 + m^2)^2} - |m|^{2(\gamma-1)} \right) m, \quad (55)$$

parametrized by $x \in (0, 1)$.

Clearly, $m = 0$ is a steady state for (55); with $\gamma = 1/2$ we interpret $m/|m| = 0$ for $m = 0$. To find nonzero steady states, we solve the algebraic equation

$$\frac{c^2 B(x)^2}{(1 + m^2)^2} - |m|^{2(\gamma-1)} = 0.$$

We distinguish the cases:

- $\gamma > 1$: The ODE (55) has three stationary points: *unstable* $m_0 = 0$ and *stable* $\pm m_s \neq 0$. Therefore, the asymptotic steady state for (55) subject to the initial datum $m^I = m^I(x)$ on $(0, 1)$ is $m_s(x)\text{sign}(m^I(x))$.

- $\gamma = 1$:

- If $c|B(x)| > 1$, then there are three stationary points, *unstable* $m_0 = 0$ and *stable* $\pm \sqrt{c|B(x)| - 1}$.
- If $c|B(x)| \leq 1$, then there is the only *stable* stationary point $m = 0$.

Thus, the solution of (55) subject to the initial datum $m^I = m^I(x)$ on $(0, 1)$ converges to the asymptotic steady state $\chi_{\{c|B(x)| > 1\}}(x)\text{sign}(m^I(x))\sqrt{c|B(x)| - 1}$.

- For $1/2 \leq \gamma < 1$ (in fact for $-1 < \gamma$, but we discard the values of $\gamma < 1/2$), the picture depends on the size of $c|B(x)|$ relative to

$$Z_\gamma := \frac{2}{\gamma + 1} \left(\frac{1 - \gamma}{1 + \gamma} \right)^{\frac{\gamma-1}{2}}. \quad (56)$$

- If $c|B(x)| > Z_\gamma$, then (55) has five stationary points, *stable* $m_0 = 0$, *unstable* $\pm m_u$ and *stable* $\pm m_s$, with $0 < m_u < m_s$.
- If $c|B(x)| = Z_\gamma$, then zero is a *stable* stationary point and there are two symmetric nonzero stationary points (attracting from $\pm\infty$ and repulsing toward zero).
- If $c|B(x)| < Z_\gamma$, then there is the only *stable* stationary point $m = 0$.

The above analysis illustrates that the structure of steady states for $1/2 \leq \gamma < 1$ is significantly richer than for $\gamma > 1$. Therefore, we may expect also in the multidimensional setting a rich structure of steady states, and, in particular, formation of complex network patterns. This hypothesis is supported by the numerical results presented in Section 4.

3.3.2 Stationary Solutions in the Multidimensional Setting

In the multidimensional setting, we are able to construct *pointwise* stationary solutions of (1)–(2). Regarding the number of possible solutions, we obtain the same picture as in the previous Section 3.3.1. However, we are not able to provide a stability analysis.

We denote the flux $u := -(I + m \otimes m)\nabla p$, so that (1) is written as

$$\nabla \cdot u = S$$

and

$$\nabla p = -(I + m \otimes m)^{-1}u = -\left(I - \frac{m \otimes m}{1 + |m|^2}\right)u. \quad (57)$$

The activation term $c^2(m \cdot \nabla p)\nabla p$ in (2) is then expressed in terms of u as

$$c^2(m \cdot \nabla p)\nabla p = c^2 \frac{m \cdot u}{1 + |m|^2} \left(I - \frac{m \otimes m}{1 + |m|^2}\right)u.$$

Therefore, stationary solutions of (1)–(2) with $D = 0$ satisfy

$$c^2 \frac{m \cdot u}{1 + |m|^2} u = \left(c^2 \frac{(m \cdot u)^2}{(1 + |m|^2)^2} + |m|^{2(\gamma-1)}\right)m. \quad (58)$$

Clearly, $m(x) = 0$ is a solution for any $u \in \mathbb{R}^d$. On the other hand, if $m(x) \neq 0$, then there exists a nonzero scalar $\beta(x) \in \mathbb{R} \setminus \{0\}$ such that $m(x) = \beta(x)u(x)$. Denoting $z := \beta(x)|u(x)|$ and inserting into (58) gives

$$\frac{c^2|u|^2}{1+z^2} = \frac{c^2|u|^2z^2}{(1+z^2)^2} + |z|^{2(\gamma-1)},$$

which further reduces to

$$c|u| = |z|^{\gamma-1}(1+z^2). \quad (59)$$

We now distinguish the cases:

- For $\gamma > 1$, the equation (59) has exactly one positive solution $z > 0$ for every $|u| > 0$.
- For $\gamma = 1$, the equation (59) has exactly one positive solution $z > 0$ for every $|u| > 1/c$ and no positive solutions for $|u| \leq 1/c$.
- For $1/2 \leq \gamma < 1$ (in fact for $-1 < \gamma$, but we discard the values of $\gamma < 1/2$), if $c|u| > Z_\gamma$ with Z_γ given by (56), there exist exactly two positive solutions $z_1, z_2 > 0$ of (59) for every $c|u| > 0$. If $c|u| = Z_\gamma$, there is one positive solution $z > 0$, and if $c|u| < Z_\gamma$, (59) has no solutions.

Let us note that in [20] stationary solutions (m_0, p_0) of (1)–(2) were considered in the case $D = 0$, $\gamma > 1$. These are constructed by fixing measurable disjoint sets $\mathcal{A}_+ \subseteq \Omega$, $\mathcal{A}_- \subseteq \Omega$ and setting

$$m_0(x) := (\chi_{\mathcal{A}_+}(x) - \chi_{\mathcal{A}_-}(x)) c^{\frac{1}{\gamma-1}} |\nabla p_0(x)|^{\frac{2-\gamma}{\gamma-1}} \nabla p_0(x), \quad (60)$$

where $p_0 \in H_0^1(\Omega)$ solves the nonlinear Poisson equation

$$-\nabla \cdot \left[\left(1 + c^{\frac{2}{\gamma-1}} |\nabla p_0(x)|^{\frac{2}{\gamma-1}} \chi_{\mathcal{A}_+ \cup \mathcal{A}_-}(x) \right) \nabla p_0(x) \right] = S, \quad (61)$$

subject to homogeneous Dirichlet boundary condition. The steady states $p_0 \in H_0^1(\Omega) \cap W_0^{1,2\gamma/(\gamma-1)}(\mathcal{A}_+ \cup \mathcal{A}_-)$ are found as the unique minimizers of the uniformly convex and coercive functional

$$\mathcal{F}_\gamma[p] := \frac{1}{2} \int_{\Omega} |\nabla p|^2 dx + c^{\frac{2}{\gamma-1}} \frac{\gamma-1}{2\gamma} \int_{\mathcal{A}_+ \cup \mathcal{A}_-} |\nabla p|^{\frac{2\gamma}{\gamma-1}} dx - \int_{\Omega} pS dx,$$

see Theorem 6 in [20]. Let us remark that the linearized stability analysis performed in Section 6.2 of [20] implies that in the case $D = 0$, $\gamma > 1$ the *linearly* stable (in the sense of Gâteaux derivative) steady states fill up the whole domain due to the necessary condition $\text{meas}(\mathcal{A}_+ \cup \mathcal{A}_-) = \text{meas}(\Omega)$. In the 1d case, the nonlinear stability analysis of Section 3.3.1 above implies that the same holds also for the (nonlinearly) stable stationary solution. On the other hand, for $\gamma = 1$ the stationary solution m_0 must vanish on the set $\{x \in \Omega; |u(x)| \leq 1/c\}$. We shall study this case below.

3.3.3 Stationary Solutions in the Multidimensional Setting for $D = 0, \gamma = 1$

In the case $\gamma = 1$, the stationary version of (2) with $D = 0$ reads

$$c^2(\nabla p_0 \otimes \nabla p_0)m_0 = m_0,$$

i.e., m_0 is either the zero vector or an eigenvector of the matrix $c^2(\nabla p_0 \otimes \nabla p_0)$ with eigenvalue 1. The spectrum of $c^2(\nabla p_0 \otimes \nabla p_0)$ consists of zero and $c^2|\nabla p_0|^2$, so that $m_0 \neq 0$ is only possible if $c^2|\nabla p_0|^2 = 1$. Therefore, for every stationary solution, there exists a measurable function $\lambda = \lambda(x)$ such that

$$m_0(x) = \lambda(x)\chi_{\{c^2|\nabla p_0|^2=1\}}(x)\nabla p_0(x)$$

and p_0 solves the highly nonlinear Poisson equation

$$-\nabla \cdot \left[\left(1 + \frac{\lambda(x)^2}{c^2} \chi_{\{c^2|\nabla p_0|^2=1\}}(x) \right) \nabla p_0 \right] = S$$

subject to the homogeneous Dirichlet boundary condition $p_0 = 0$ on $\partial\Omega$.

A simple consideration suggests that *stable* stationary solutions of (1)–(2) with $D = 0$ should be constructed as

$$-\nabla \cdot [(1 + a(x)^2)\nabla p_0] = S, \quad p_0 \in H_0^1(\Omega), \quad (62)$$

$$c^2|\nabla p_0(x)|^2 \leq 1, \quad \text{a.e. on } \Omega, \quad (63)$$

$$a(x)^2 [c^2|\nabla p_0(x)|^2 - 1] = 0, \quad \text{a.e. on } \Omega, \quad (64)$$

for some measurable function $a^2 = a(x)^2$ on Ω which is the Lagrange multiplier for the condition (63). This condition follows from the nonpositivity of the eigenvalues of the matrix $c^2(\nabla p_0 \otimes \nabla p_0) - I$, which is heuristically a necessary condition for linearized stability of the stationary solution of (1)–(2) with $D = 0$. The function $\lambda = \lambda(x)$ can be chosen as $\lambda(x) := ca(x)$.

Note that this construction is compatible with the result of Section 3.3.2: either $m = 0$, then $|u| = |\nabla p| \leq 1/c$, or $m \neq 0$, then $|\lambda| > 0$ and

$$|u| = (1 + \lambda^2|\nabla p|^2)|\nabla p| = \left(1 + \frac{\lambda^2}{c^2} \right) \frac{1}{c} > \frac{1}{c}.$$

We claim that solutions of (62)–(64) are minimizers of the energy functional

$$\mathcal{J}[p] := \int_{\Omega} \left(\frac{|\nabla p|^2}{2} - Sp \right) dx \quad (65)$$

on the set $\mathcal{M} := \{p \in H_0^1(\Omega), c^2|\nabla p|^2 \leq 1 \text{ a.e. on } \Omega\}$.

Lemma 11 Let $S \in L^2(\Omega)$. There exists a unique minimizer of the functional (65) on the set \mathcal{M} . It is the unique weak solution of the problem (62)–(64) with homogeneous Dirichlet boundary conditions on Ω and with $a \in L^2(\Omega)$.

Proof The functional \mathcal{J} is convex and, due to the Poincaré inequality, coercive on $H_0^1(\Omega)$. Therefore, a unique minimizer $p_0 \in H_0^1(\Omega)$ exists on the closed, convex set \mathcal{M} . Clearly, (62)–(64) is the Euler–Lagrange system corresponding to this constrained minimization problem, so that p_0 is its weak solution. Moreover, using p_0 as a test function and an application of the Poincaré inequality yields

$$\begin{aligned} \int_{\Omega} (1 + a^2) |\nabla p_0|^2 dx &= \int_{\Omega} S p_0 dx \\ &\leq \int_{\Omega} |\nabla p_0|^2 dx + C \int_{\Omega} S^2 dx. \end{aligned}$$

With (64), we have then

$$\int_{\Omega} a^2 dx = c^2 \int_{\Omega} a^2 \nabla |p_0|^2 dx \leq c^2 C \int_{\Omega} S^2 dx,$$

so that $a \in L^2(\Omega)$.

Next, we prove that any weak solution $p_0 \in H_0^1(\Omega)$, $a^2 \in L^1(\Omega)$ of (62)–(64) is a minimizer of (65) on the set \mathcal{M} . Indeed, we consider any $q \in \mathcal{M}$ and use $(p_0 - q)$ as a test function for (62),

$$\int_{\Omega} (1 + a^2) \nabla p_0 \cdot \nabla (p_0 - q) dx = \int_{\Omega} S(p_0 - q) dx.$$

The Cauchy–Schwarz inequality for the term $\nabla p_0 \cdot \nabla q$ gives

$$\frac{1}{2} \int_{\Omega} (1 + a^2) |\nabla p_0|^2 dx \leq \frac{1}{2} \int_{\Omega} (1 + a^2) |\nabla q|^2 dx + \int_{\Omega} (p_0 - q) S dx.$$

Moreover, (64) gives $a^2 |\nabla p_0|^2 = a^2 / c^2$, and with $|\nabla q| \leq 1/c^2$ we have

$$\int_{\Omega} \left(\frac{|\nabla p_0|^2}{2} - p_0 S \right) dx + \frac{1}{2c^2} \int a^2 dx \leq \int_{\Omega} \frac{|\nabla q|^2}{2} - q S dx + \frac{1}{2c^2} \int a^2 dx,$$

so that $\mathcal{J}[p_0] \leq \mathcal{J}[q]$.

Finally, let $p_i \in H_0^1(\Omega)$, $a_i \in L^2(\Omega)$, $i = 1, 2$, be two weak solutions of (62)–(64). We take the difference of (62) for p_1 and p_2 and test by $(p_1 - p_2)$:

$$\int_{\Omega} [(1 + a_1^2) \nabla p_1 - (1 + a_2^2) \nabla p_2] \cdot (\nabla p_1 - \nabla p_2) dx = 0.$$

We use the Cauchy–Schwarz inequality for

$$\begin{aligned} \int_{\Omega} (a_1^2 + a_2^2)(\nabla p_1 \cdot \nabla p_2) dx &\leq \frac{1}{2} \int_{\Omega} (a_1^2 + a_2^2)|\nabla p_1|^2 dx + \frac{1}{2} \int_{\Omega} (a_1^2 + a_2^2)|\nabla p_2|^2 dx \\ &\leq \frac{1}{c^2} \int_{\Omega} (a_1^2 + a_2^2), \end{aligned}$$

where the second inequality comes from (63). Consequently, we have

$$\int_{\Omega} |\nabla p_1 - \nabla p_2|^2 dx + \int_{\Omega} a_1^2 |\nabla p_1|^2 + a_2^2 |\nabla p_2|^2 \leq \frac{1}{c^2} \int_{\Omega} (a_1^2 + a_2^2).$$

Finally, using (64) we obtain

$$\int_{\Omega} |\nabla p_1 - \nabla p_2|^2 dx \leq 0$$

and conclude that $p_1 = p_2$ a.e. on Ω . \square

Remark 7 The gradient constrained variational problem (65) was studied in [9] as a model for twisting of an elastic–plastic cylindrical bar. There it was shown that the unique solution has $C^{1,1}$ -regularity in Ω ; see also [39, 40].

Remark 8 Solutions of (62)–(64) subject to periodic boundary conditions on $\partial\Omega$ can also be constructed via the following penalty approximation,

$$-\nabla \cdot \left[\left(1 + \frac{(|\nabla p_\varepsilon|^2 - 1/c^2)_+}{\varepsilon} \right) \nabla p_\varepsilon \right] = S, \quad p_\varepsilon \in \bar{H}^1(\Omega), \quad (66)$$

with $\varepsilon > 0$, where $A_+ := \max(A, 0)$ denotes the positive part of A and $\bar{H}^1(\Omega) = \{u \in H_{\text{per}}^1(\Omega), \int_{\Omega} u dx = 0\}$. Here $H_{\text{per}}^1(\Omega)$ denotes the space of $H_{\text{loc}}^1(\mathbb{R}^d)$ -functions with $(0, 1)^d$ -periodicity.

This approach was considered in [22], where it was proven that for each $S \in L^2(\Omega)$ with $\int_{\Omega} S(x) dx = 0$ a unique weak solution $p_\varepsilon \in \bar{H}^1(\Omega)$ of (66) exists. Moreover, the sequence $(p_\varepsilon)_{\varepsilon>0}$ converges to a weak solution of the system (62)–(64) as $\varepsilon \rightarrow 0$; see Theorems 2 and 3 of [22].

4 Numerical Methods

The network model has complex dynamics, which lead to interesting patterns. To demonstrate this numerically, let \mathcal{T}_h be a shape regular, quasi-uniform triangulation of $\Omega \subset \mathbb{R}^2$, cf. [7, Chapter 4]. For the approximation of the conductance and the pressure, let us introduce the space of continuous and piecewise linear functions

$$V_h := \{\varphi \in C^0(\overline{\Omega}) : \varphi|_T \in \mathcal{P}_1(T) \text{ for } T \in \mathcal{T}_h\}$$

associated with \mathcal{T}_h . As an auxiliary space, we also consider

$$W_h := \{\chi \in L^2(\Omega) : \chi|_T \in \mathcal{P}_0(T) \text{ for } T \in \mathcal{T}_h\},$$

which consists of piecewise constant functions. Here, $\mathcal{P}_k(T)$ denotes the set of polynomials of degree at most $k \in \mathbb{N}_0$. Choosing the Lagrangian basis $\{\Phi_i : i = 1, \dots, M\}$ of V_h associated with the set of vertices $\{x_i : i = 1, \dots, M\}$ in \mathcal{T}_h , we may identify each function $\varphi = \sum_{i=1}^M \varphi_i \Phi_i \in V_h$ with its coefficient vector $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_M)^\top \in \mathbb{R}^M$. Similarly, we denote by $\{\chi_j = \chi_{T_j} : j = 1, \dots, N\}$ the canonical basis of W_h .

Since, among other things, we are interested in the influence of the diffusion parameter D on the network formation, we want to construct an approximation scheme which does not introduce numerical diffusion. Therefore, we will employ a mass lumping strategy. For $\mathbf{M}^l \in \mathbb{R}^{M \times M}$, defined by $\mathbf{M}_{i,j}^l := 0$ if $i \neq j$ and

$$\mathbf{M}_{i,i}^l := \omega_i := \int_{\Omega} \Phi_i(x) dx = \frac{1}{3} |\text{supp}(\Phi_i)|,$$

let us introduce a scalar product along with its induced norm by

$$(\varphi, \psi)_{M^l} := \sum_{i=1}^M \omega_i \varphi_i \psi_i, \quad \varphi, \psi \in V_h, \quad \|\varphi\|_{M^l} := (\varphi, \varphi)_{M^l}^{1/2}.$$

In order to discretize the activation term without introducing numerical diffusion, we also employ a local averaging procedure, cf. [10]. To this end, we define a linear operator $A_h : W_h \rightarrow V_h$, where $A_h \chi \in V_h$ is the solution to the variational problem

$$(A_h \chi, \varphi)_{M^l} = (\chi, \varphi)_{L^2(\Omega)} \quad \text{for all } \varphi \in V_h.$$

Writing $A_h \chi = \sum_{i=1}^M \mathbf{a}_i \Phi_i$, we see that

$$\mathbf{a}_i = \frac{1}{|\text{supp}(\Phi_i)|} \sum_{T \subset \text{supp}(\Phi_i)} \chi_T |T|. \quad (67)$$

Using (67) and Jensen's inequality, we obtain for $\chi \in W_h$

$$\|A_h \chi\|_{M^l}^2 = \sum_{i=1}^M \omega_i \mathbf{a}_i^2 \leq \sum_{i=1}^M \frac{\omega_i}{|\text{supp}(\Phi_i)|} \sum_{T \subset \text{supp}(\Phi_i)} |T| |\chi_T|^2 = \frac{1}{3} \sum_{j=1}^N |T_j| n(T_j) |\chi_{T_j}|^2.$$

Here, $n(T_j) \in \mathbb{N}$ is the number of patches $\text{supp}(\Phi_i)$ containing T_j , i.e., in dimension two, we have $n(T_j) = 3$. Thus, A_h is L^2 -stable, i.e.

$$\|A_h \chi\|_{M^l} \leq \|\chi\|_{L^2(\Omega)} \quad \text{for all } \chi \in W_h. \quad (68)$$

In slight abuse of notation, we will apply A_h to vector-valued quantities meaning that we apply A_h component wise.

We are now in the position to define a discrete energy functional

$$E_h(m_h) := \frac{D^2}{2} \|\nabla m_h\|_{L^2(\Omega)}^2 + \frac{c^2}{2} \left(\|m_h \cdot A_h \nabla p_h\|_{M^l}^2 + r \|\nabla p_h\|_{L^2(\Omega)}^2 \right) + \frac{1}{2\gamma} \||m_h|^\gamma\|_{M^l}^2, \quad (69)$$

where $m_h \in V_h \times V_h$ and $p_h = p_h[m_h] \in V_h$ is a solution to the Kirchhoff law

$$a_{K,h}(p_h, \varphi) = \ell_{S,G}(\varphi) \quad \text{for all } \varphi \in V_h, \quad (70)$$

with normalization $\int_\Omega p_h \, dx = 0$ and bilinear form $a_{K,h} : V_h \times V_h \rightarrow \mathbb{R}$ and linear form $\ell_{S,G} : V_h \rightarrow \mathbb{R}$ defined by

$$\begin{aligned} a_{K,h}(p_h, \varphi) &:= (r \nabla p_h, \nabla \varphi)_{L^2(\Omega)} + (m_h \cdot A_h \nabla p_h, m_h \cdot A_h \nabla \varphi)_{M^l}, \\ \ell_{S,G}(\varphi) &:= (S, \varphi)_{L^2(\Omega)} + (G, \varphi)_{L^2(\partial\Omega)}. \end{aligned}$$

The functions S and G , which models Neumann boundary conditions for the pressure, should satisfy the compatibility condition $\ell_{S,G}(1) = 0$. In view of the Lax–Milgram lemma and the Poincaré inequality, (70) has a unique solution $p_h \in V_h$ satisfying $\int_\Omega p_h \, dx = 0$ for each fixed $m_h \in V_h \times V_h$. With similar computations as in Section 2.3, we obtain the following gradient flow for the minimization of E_h

$$\begin{aligned} \frac{d}{dt} (m_h, \Phi_i)_{M^l} &= -D^2 (\nabla m_h, \nabla \Phi_i)_{L^2(\Omega)} \\ &\quad + c^2 (A_h \nabla p_h \otimes A_h \nabla p_h m_h, \Phi_i)_{M^l} - (|m_h|^{2(\gamma-1)} m_h, \Phi_i)_{M^l}. \end{aligned}$$

for $i = 1, \dots, M$ and $\gamma > 1/2$. Equivalently, for the component vector $\mathbf{m} \in \mathbb{R}^{M \times 2}$ of m_h defined row-wise by $\mathbf{m}_i = (\mathbf{m}_{1,i}, \mathbf{m}_{2,i}) = (m_{h,1}(x_i), m_{h,2}(x_i)) \in \mathbb{R}^{1 \times 2}$, we obtain the ODE system

$$\frac{d}{dt} \mathbf{m}_i = -D^2 ((\mathbf{M}^l)^{-1} \mathbf{K} \mathbf{m})_i + c^2 \mathbf{m}_i (A_h \nabla p_h)_i \otimes (A_h \nabla p_h)_i - |\mathbf{m}_i|^{2(\gamma-1)} \mathbf{m}_i, \quad (71)$$

with initial condition $\mathbf{m}_i(0) = (m_{0,1}(x_i), m_{0,2}(x_i))$, $i = 1, \dots, M$, for sufficiently regular m_0 . Here, we denote $(A_h \nabla p_h)_i = (A_h \nabla p_h)(x_i) \in \mathbb{R}^2$, and the stiffness matrix $\mathbf{K} \in \mathbb{R}^{M \times M}$ is defined by

$$\mathbf{K}_{i,j} = (\nabla \Phi_i, \nabla \Phi_j)_{L^2(\Omega)}, \quad i, j \in \{1, \dots, M\}.$$

The case $\gamma = 1/2$ needs some modifications, see (48) and (78) below.

Remark 9 For $D = 0$, the ODE system (71) decouples, and the support of m_h cannot grow. We note that we could replace A_h by any other (averaging) operator which

maps gradients of the pressure, which are piecewise constants, to functions in V_h as long as (70) is uniquely solvable. A key observation is that the activation term in (69) equals $c^2 a_{K,h}(p_h, p_h)/2$, i.e., changing the activation term in the gradient flow, resp., the energy, necessarily requires a modification of the Poisson equation in order to guarantee energy decrease along solutions of the gradient flow. Another discretization strategy avoiding numerical diffusion can be built upon mixed finite elements, see [22].

The three terms on the right-hand side of (71) possess very different dynamics if D is small and c is large. This motivates the use of splitting methods to discretize the ODE system (71) in time. We discuss each term separately in the following.

4.1 Activation

For $c|A_h \nabla p_h| \gg 1$, the ODE

$$\partial_t \mathbf{m}_i = c^2 \mathbf{m}_i (A_h \nabla p_h)_i \otimes (A_h \nabla p_h)_i, \quad \mathbf{m}_i(0) = (m_{0,1}(x_i), m_{0,2}(x_i)),$$

governing the activation is stiff and implicit methods might be preferred. However, the resulting nonlinear equations are expensive to solve, and, therefore, we will employ the explicit Euler method

$$\mathbf{m}_i^{k+1} = \mathbf{m}_i^k + \delta t c^2 \mathbf{m}_i^k (A_h \nabla p_h)_i \otimes (A_h \nabla p_h)_i. \quad (72)$$

Here, we denote by δt a time step size and by \mathbf{m}_i^k an approximation to $\mathbf{m}_i(t^k)$ with $t^k = k\delta t$. To obtain stable solutions, it is important to choose a suitable step size δt . Since (72) amounts to a gradient descent step for the functional

$$h(m_h) := \frac{c^2}{2} (r \|\nabla p_h\|_{L^2(\Omega)}^2 + \|m_h \cdot A_h \nabla p_h\|_{M^l}^2) \quad (73)$$

a reasonable time step size should be related to the Lipschitz constant of ∇h ; cf. Section 4.5 below. Using the identity $h(m_h) = \frac{c^2}{2} a_{K,h}(p_h, p_h)$, we see that

$$\delta_m^2 h[m_h](v, w) = c^2 (a_{K,h}(\delta_m p_h[w], \delta_m p_h[v]) - (v \cdot A_h \nabla p_h, w \cdot A_h \nabla p_h)_{M^l}), \quad (74)$$

where $v, w \in V_h \times V_h$, and $\delta_m p_h[v] \in V_h$ denotes the directional derivative of p_h with respect to m_h in direction v , i.e., $\delta_m p_h[v]$ is the solution with zero mean to

$$a_{K,h}(\delta_m p_h[v], \varphi) = -(v \cdot A_h \nabla p_h, m_h \cdot A_h \nabla \varphi)_{M^l} - (m_h \cdot A_h \nabla p_h, v \cdot A_h \nabla \varphi)_{M^l} \quad (75)$$

for all $\varphi \in V_h$. To estimate $\delta_m^2 h[m_h]$, we first derive a bound for $\nabla \delta_m p_h[v]$. Testing (75) with $\delta_m p_h[v]$ and applying the Cauchy–Schwarz inequality yields

$$a_{K,h}(\delta_m p_h[v], \delta_m p_h[v]) \leq \bar{C}(r \|A_h \nabla \delta_m p_h[v]\|_{M^l}^2 + \|m_h \cdot A_h \nabla \delta_m p_h[v]\|_{M^l}^2)^{1/2} \|v\|_{M^l},$$

where

$$\bar{C} = \sup_{i=1,\dots,M} \left(|m_h(x_i) \cdot (A_h \nabla p_h)(x_i)|^2 / r + |(A_h \nabla p_h)(x_i)|^2 \right)^{1/2}.$$

Using L^2 -stability of A_h (68), we further obtain

$$a_{K,h}(\delta_m p_h[v], \delta_m p_h[v])^{1/2} \leq \bar{C} \|v\|_{M^l}. \quad (76)$$

Using (74), (76) and the Cauchy–Schwarz inequality, we conclude that

$$|\delta_m^2 h[m_h](v, w)| \leq \text{Lip}(\nabla h(m_h)) \|v\|_{M^l} \|w\|_{M^l},$$

where we define an estimate for the local Lipschitz constant of ∇h by

$$\text{Lip}(\nabla h(m_h)) = c^2 \left(\bar{C}^2 + \sup_{i=1,\dots,M} |(A_h \nabla p_h)(x_i)|^2 \right). \quad (77)$$

4.2 Relaxation

For $\gamma > 1/2$, we consider the ODE

$$\partial_t \mathbf{m}_i = -|\mathbf{m}_i|^{2(\gamma-1)} \mathbf{m}_i,$$

and for $\gamma = 1/2$, we consider the subgradient differential inclusion

$$-\partial_t \mathbf{m}_i \in \partial_{|\cdot|}(\mathbf{m}_i). \quad (78)$$

Here, $\partial_{|\cdot|}$ is the subdifferential of the Euclidean norm. Let us discuss the case $\gamma = 1/2$ in more detail. In [22], a regularized version has been considered, i.e. $|m_h|$ in (69) has been approximated by a pseudo-Huber function

$$|m_h|_\rho := \sqrt{m_{h,1}^2 + m_{h,2}^2 + \rho^2}$$

with regularization parameter $\rho > 0$ leading to the ODE

$$\partial_t \mathbf{m}_i = -\frac{1}{|\mathbf{m}_i|_\rho} \mathbf{m}_i.$$

Applying the explicit Euler method with step size δt yields the update formula

$$\mathbf{m}_i^{k+1} := \tilde{\lambda}_i^k \mathbf{m}_i^k \quad \text{with} \quad \tilde{\lambda}_i^k := (1 - \frac{\delta t}{|\mathbf{m}_i^k|_\rho}).$$

The condition $\tilde{\lambda}_i^k \geq 0$ requires the time step restriction $\delta t \leq \rho$ in the limit $|\mathbf{m}_i^k| \rightarrow 0$. If this condition is not satisfied, i.e., $\tilde{\lambda}_i^k < 0$, \mathbf{m}_i^{k+1} will point into the opposite direction of \mathbf{m}_i^k . This is a likely explanation of the oscillatory behavior in the simulation in [1, 22]. Note that, for $\rho = 0$ and $|\mathbf{m}_i^k| \approx 0$, the explicit Euler method will break down if $\tilde{\lambda}_i^k \geq 0$ is required. Thus, we suggest to use an implicit method.

Let us revisit the implicit Euler method, which is determined by the formula

$$m_h^{k+1} := \operatorname{argmin}_{v_h \in V_h^2} \frac{1}{2\delta t} \|v_h - m_h^k\|_{M_l}^2 + \frac{1}{2\gamma} \||v_h|^\gamma\|_{M_l}^2.$$

The solution of the minimization problem is determined by the coefficient update

$$\mathbf{m}_i^{k+1} := \lambda_i^k \mathbf{m}_i^k. \quad (79)$$

If $\gamma = 1/2$, we have the soft shrinkage formula

$$\lambda_i^k := \begin{cases} \frac{(|\mathbf{m}_i^k| - \delta t)_+}{|\mathbf{m}_i^k|} & , |\mathbf{m}_i^k| > 0, \\ 0 & , |\mathbf{m}_i^k| = 0. \end{cases}$$

If $\gamma > 1/2$, we set $\lambda_i^k = 0$ if $|\mathbf{m}_i^k| = 0$, and otherwise we let $\lambda_i^k \in (0, 1)$ be defined as the unique solution to

$$\lambda + \delta t |\mathbf{m}_i^k|^{2(\gamma-1)} \lambda^{2\gamma-1} = 1,$$

which can be computed using bisection.

4.3 Diffusion

The semi-discrete heat equation

$$\partial_t \mathbf{M}^l \mathbf{m} = -D^2 \mathbf{K} \mathbf{m}, \quad \mathbf{m}_i(0) = (m_{0,1}(x_i), m_{0,2}(x_i)),$$

can be solved exactly

$$\mathbf{m}(t + \delta t) = \exp(-\delta t D^2 (\mathbf{M}^l)^{-1} \mathbf{K}) \mathbf{m}(t), \quad (80)$$

and efficiently via Krylov subspace methods, see [6] for an introduction and further references. We are interested in the case $D^2 \ll 1$. While $\delta t D^2 \ll 1$ is favorable in view of an efficient application of the matrix exponential by Krylov subspace methods, one should take a sufficiently fine discretization such that $\delta t D^2 / h^2$ is orders of magnitude larger than machine precision. If $\delta t D^2 / h^2$ is comparable to machine precision, the diffusion process might introduce additional randomness which in turn is strongly enhanced by the activation term if $c|A_h \nabla p_h| \gg 1$. As a consequence, the computed solution might be unsymmetric. Let us also refer to the discussion in [22].

4.4 Algorithm 1: Splitting

Network patterns will occur, if D is small and c is large, i.e., if diffusion is dominated by activation. This immediately leads to very different dynamics of the different terms in the gradient flow (71). On the one hand, as observed in [22], at an early stage in the evolution $Lip(\nabla h(m_h))$ as defined in (77) is very large. This in turn leads to small time steps $\delta t \approx 1/Lip(\nabla h(m_h))$ if the dynamics should be computed accurately. On the other hand, this implies that $D^2 \delta t \ll 1$ and prohibitively fine grids should be used in order to resolve diffusion well. To account for these different dynamics, we propose the following algorithm.

Initialization. Choose a step size δt such that $\delta t D^2 / h^2 \approx 1$. Assume, we have given an iterate m_h^k at a time instance t^k . Set $\tilde{m}_h = m_h^k$, $t_{loc} = t^k$, and fix $\xi \in (0, 1)$.

1. Step: Activation. Define $\delta t_{loc} := \xi / Lip(\nabla h(\tilde{m}_h))$. If $t_{loc} + \delta t_{loc} \geq t^k + \delta t$, set $\delta t_{loc} := \max(0, t^k + \delta t - \delta t_{loc})$. Take an explicit Euler step with step size δt_{loc} according to (72) to update \tilde{m}_h . Set $t_{loc} := t_{loc} + \delta t_{loc}$. If $t_{loc} = t^k + \delta t$, break and go to step 2; otherwise repeat step 1.

2. Step: Relaxation. Update \tilde{m}_h according to (79).

3. Step: Diffusion. Update \tilde{m}_h according to (80) and set $m_h^{k+1} := \tilde{m}_h$.

4. Step: Check energy decrease. If $E_h(m_h^{k+1}) > E_h(m_h^k)$ return m_h^k . If $E_h(m_h^{k+1}) < E_h(m_h^k)$, set $\tilde{m}_h := m_h^{k+1}$ and increase k by one, go to step 1.

Note that energy decrease is not guaranteed here. In view of the results presented in [22], we expect that $Lip(\nabla h(\tilde{m}_h))$ decreases during the evolution. If $\delta t_{loc} = \delta t$, then the above splitting algorithm is similar to the forward–backward splitting algorithm described in the next section, cf. Remark 10. For the forward–backward splitting algorithm, there is a simple criterion regarding the choice of the time step size guaranteeing energy decrease.

4.5 Algorithm 2: Forward–Backward Splitting

Let us rewrite the energy as the sum of two functionals

$$E_h(m_h) = g(m_h) + h(m_h),$$

where h as defined in (73) is differentiable with Lipschitz continuous gradient, see (77), but not convex in general. The strictly convex functional g is defined as

$$g(v_h) := \frac{D^2}{2} \|\nabla v_h\|_{L^2(\Omega)}^2 + \frac{1}{2\gamma} \||v_h|^\gamma\|_{M'}^2, \quad v_h \in V_h \times V_h.$$

The forward–backward splitting algorithm reads as [37]

$$m_h^{k+1} := \text{prox}_{\delta t_k g}(m_h^k - \delta t_k \nabla h(m_h^k)), \quad (81)$$

where δt_k corresponds to a variable time step size. Recall that the proximal operator of a convex, proper and lower-semicontinuous functional Γ is defined by

$$\hat{x} := \text{prox}_\Gamma(x) := \arg\min_z \frac{1}{2} \|z - x\|_2^2 + \Gamma(z).$$

Therefore, evaluation of $\text{prox}_{\delta t_k g}(f^k)$ amounts to computing the minimizer of

$$\frac{1}{2\delta t_k} \|v_h - f^k\|_{M'}^2 + \frac{D^2}{2} \|\nabla v_h\|_{L^2(\Omega)}^2 + \frac{1}{2\gamma} \||v_h|^\gamma\|_{M'}^2 \rightarrow \min_{v_h \in V_h \times V_h},$$

here $f^k = m_h^k - \delta t_k \nabla h(m_h^k)$. This minimization problem can be regarded as a discrete counterpart of (32). The solution of this problem can, for instance, be obtained using the Douglas–Rachford splitting [15, 37]. Setting

$$\tilde{h}(v_h) := \frac{1}{2\delta t_k} \|v_h - f^k\|_{M'}^2 + \frac{D^2}{2} \|\nabla v_h\|_{L^2(\Omega)}^2 \quad \text{and} \quad \tilde{g}(v_h) := \frac{1}{2\gamma} \||v_h|^\gamma\|_{M'}^2,$$

the Douglas–Rachford iteration computes a fixed point of the operator

$$T := \text{prox}_{\delta t_k \tilde{h}}(2\text{prox}_{\delta t_k \tilde{g}} - I) - \text{prox}_{\delta t_k \tilde{g}} + I,$$

which can be obtained via the following algorithm.

Initialization. Choose $\tilde{m}_h^0 \in V_h \times V_h$, $v_h^0 \in V_h \times V_h$.

Iteration. For $r = 0, 1, \dots$ do updates until $\|v_h^{r+1} - v_h^r\|_{M'} < tol$

$$\begin{aligned} v_h^{r+1} &:= \text{prox}_{\delta t_k \tilde{h}}(2\tilde{m}_h^r - v_h^r) + v_h^r - \tilde{m}_h^r, \\ \tilde{m}_h^{r+1} &:= \text{prox}_{\delta t_k \tilde{g}}(v_h^{r+1}). \end{aligned}$$

The fixed-point iteration $v_h^{r+1} = T(v_h^r)$ converges to a unique \hat{v}_h [15, 37], and

$$m_h^{k+1} := \text{prox}_{\delta t_k g}(m_h^k - \delta t_k \nabla h(m_h^k)) = \text{prox}_{\delta t_k \tilde{g}}(\hat{v}_h),$$

which can be computed using (79). The definition of $\text{prox}_{\delta t_k g}$ yields

$$\frac{1}{2\delta t_k} \|m_h^{k+1} - m_h^k + \delta t_k \nabla h(m_h^k)\|_{M^l}^2 + g(m_h^{k+1}) \leq \frac{1}{2\delta t_k} \|\delta t_k \nabla h(m_h^k)\|_{M^l}^2 + g(m_h^k).$$

For L_k denoting an upper bound on the Lipschitz constant of the mapping $t \mapsto \nabla h(m_h^k + t(m_h^{k+1} - m_h^k))$, $0 \leq t \leq 1$, the mean value theorem asserts that

$$h(m_h^{k+1}) - h(m_h^k) - \frac{L_k}{2} \|m_h^{k+1} - m_h^k\|_{M^l}^2 \leq (\nabla h(m_h^k), m_h^{k+1} - m_h^k)_{M^l}.$$

Since furthermore,

$$\begin{aligned} & \|m_h^{k+1} - m_h^k + \delta t_k \nabla h(m_h^k)\|_{M^l}^2 - \|\delta t_k \nabla h(m_h^k)\|_{M^l}^2 \\ &= \|m_h^{k+1} - m_h^k\|_{M^l}^2 + 2(\nabla h(m_h^k), m_h^{k+1} - m_h^k)_{M^l}, \end{aligned}$$

we see that

$$\frac{1}{2} \left(\frac{1}{\delta t_k} - L_k \right) \|m_h^{k+1} - m_h^k\|_{M^l}^2 + E_h(m_h^{k+1}) \leq E_h(m_h^k). \quad (82)$$

If $L_k \leq \text{Lip}(\nabla h(m_h^k))$ and $\delta t_k < 1/\text{Lip}(\nabla h(m_h^k))$, the forward–backward splitting iteration decreases the energy functional. Convergence of forward–backward splitting methods for convex functionals has been investigated, e.g., in [37], for nonconvex functionals, which is the case considered here, we refer to [2]. We have, however, not verified the assumptions of [2]. It thus remains open to prove convergence of the forward–backward splitting algorithm for the problem considered here.

Remark 10 As argued in [2], the forward–backward splitting can be computed inexactly still providing energy decrease. If we initialize the Douglas–Rachford algorithm with $\tilde{m}_h^0 = v_h^0 = m_h^k$, and perform only one iteration, this corresponds to Algorithm 1, if we replace the application of the matrix exponential of the Laplace operator by an implicit Euler step and if $\delta t_{loc} = \delta t = \delta t_k$. The case $\delta t_{loc} \ll \delta t$ is, however, different.

5 Numerical Examples

The system (1)–(2) contains the parameters D , γ , and c , which strongly influence network formation. This will be demonstrated by means of numerical examples below. As a stopping criterion for our algorithms, we define $\delta m_h^k \leq 5 \times 10^{-10}$, where

$$\delta m_h^k := \frac{1}{\delta t_{k-1}} \|m_h^k - m_h^{k-1}\|_{L^2(\Omega)}, \quad k \geq 1.$$

Due to slow convergence of the gradient flow, we will show in most of the examples below transient states of the gradient flow. The corresponding change in energy is denoted by $\delta E_h^k := (E_h(m_h^k) - E_h(m_h^{k-1})) / \delta t_{k-1}$ with E_h defined in (69). To measure the concentration of the resulting structures, let us introduce the sparsity measure

$$s_k := \|m_h^k\|_{L^2(\Omega)} / \|m_h^k\|_{L^1(\Omega)}.$$

Furthermore, we denote by $u_h^k := -(rI + m_h^k \otimes m_h^k) \nabla p_h^k$ the flow along the network. In all our examples, we use Algorithm 1 with $\delta t = 1/2$ and we set the background permeability $r = 1/10$, and the activation parameter $c = 200$. If Algorithm 1 breaks but $\delta m_h^k > 5 \times 10^{-10}$, we proceed using Algorithm 2, i.e., (81).

5.1 Example 1: One Source, Two Sinks

We consider the transport of a quantity from one source to two sinks. We choose $\Omega = (0, 1) \times (-1/2, 1/2)$ and a uniform triangulation with $M = 93\,025$ vertices, i.e., $h \approx 2.3 \times 10^{-3}$. We remark that this mesh is not symmetric with respect to the axis $x = 0$. As initial datum, we set $m_1^I = 0$ and

$$m_1^I(x, y) = e^{-10^4 y^2 - 50(x - \frac{1}{10})^2}.$$

The source terms are given by $G = 0$ and S is defined as

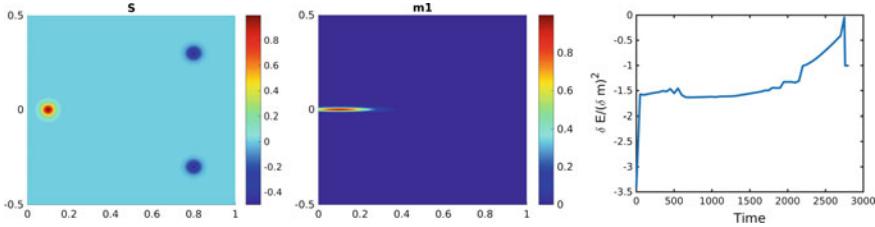


Fig. 3 Example 1: Internal sources and sinks (left) and m_1^I (middle). Evolution of $\delta E_h^k / (\delta m^k)^2$ for $\gamma = 0.9$ and $D = 0.0025$ (right).

Table 2 Example 1: Stationarity and sparsity measures for different values of γ and $D = 0.0025$.

γ	0.5	0.6	0.75	0.8	0.9	1
t^k	7246	5296	6061	5992	2810	606
δt^k	0.5	0.5	0.5	0.0046	0.0025	0.5
E^k	0.334	0.356	0.396	0.400	0.427	0.462
δm^k	3×10^{-4}	5×10^{-4}	3×10^{-4}	3×10^{-5}	3×10^{-5}	5×10^{-10}
s_k	5.68	5.53	4.91	5.01	4.36	2.28

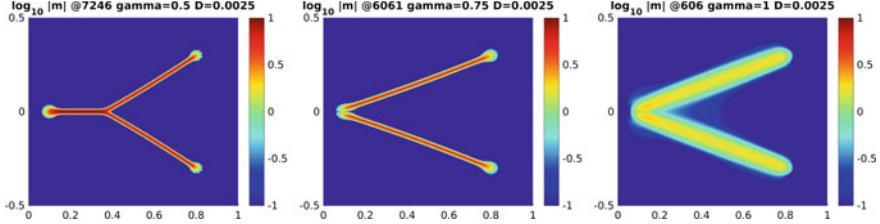


Fig. 4 Example 1: $\log_{10}(|m|)$ for $D = 0.0025$ and $\gamma \in \{0.5, 0.75, 1\}$.

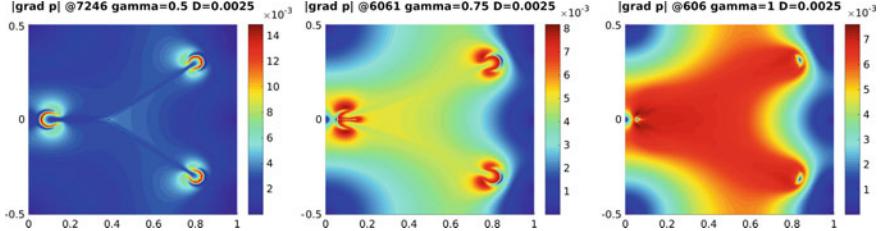


Fig. 5 Example 1: $|\nabla p|$ for $D = 0.0025$ and $\gamma \in \{0.5, 0.75, 1\}$.

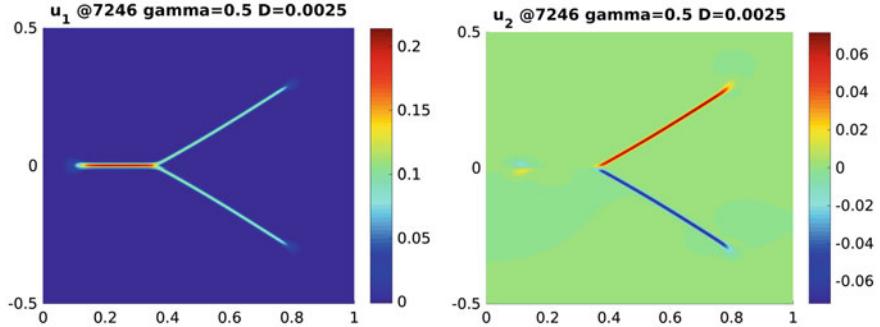


Fig. 6 Example 1: Transient state of u_1 and u_2 for $D = 0.0025$ and $\gamma = 0.5$.

$$S(x, y) = e^{-1000(y^2 + (x - \frac{1}{10})^2)} - \frac{1}{2}(e^{-1000((y - \frac{3}{10})^2 + (x - \frac{4}{5})^2)} + e^{-1000((y + \frac{3}{10})^2 + (x - \frac{4}{5})^2)}) - C$$

with $C \in \mathbb{R}$ such that $\ell_{S,G}(1) = 0$, see Figure 3. We consider the case $D = 0.0025$ and $\gamma \in \{0.5, 0.6, 0.75, 0.8, 0.9, 1\}$. The parameter γ is used in the relaxation term to model costs of maintaining the network. For $\gamma = 1$, Algorithm 1 converged.

For $\gamma = 0.9$, Algorithm 1 broke down at $t^k = 2758.5$; cf. Figure 3. For $t^k = 2758$, $E_h^k = 0.442583$, and we initialized Algorithm 2 with the corresponding iterate m^k . Algorithm 2 decreased the energy further down to $E_h^k = 0.426523$ for $t^k = 2810.405$ and then stopped due to another increase in energy. At this point, $\delta m^k \approx 10^{-5}$, $\delta E_h^k \approx 10^{-10}$ and $\delta t_k \approx 10^{-3}$. We explain this energy increase by the fact that we have to specify some tolerances in our algorithms. In order to implement (79) for

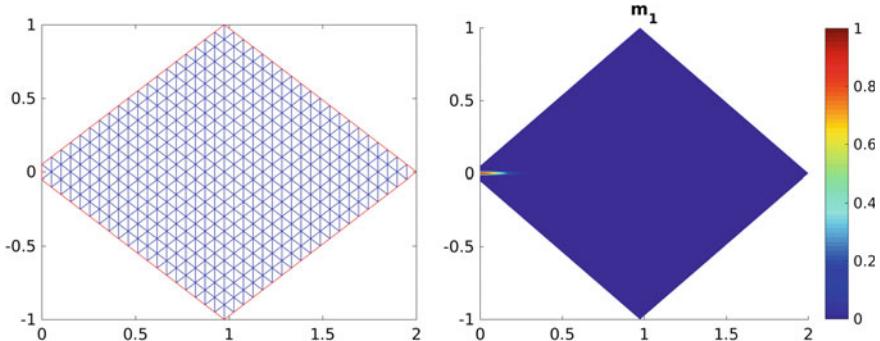


Fig. 7 Example 2: Triangulation \mathcal{T}_h of Ω with $M = 440$ vertices (left), and first component of initial datum m^I (right).

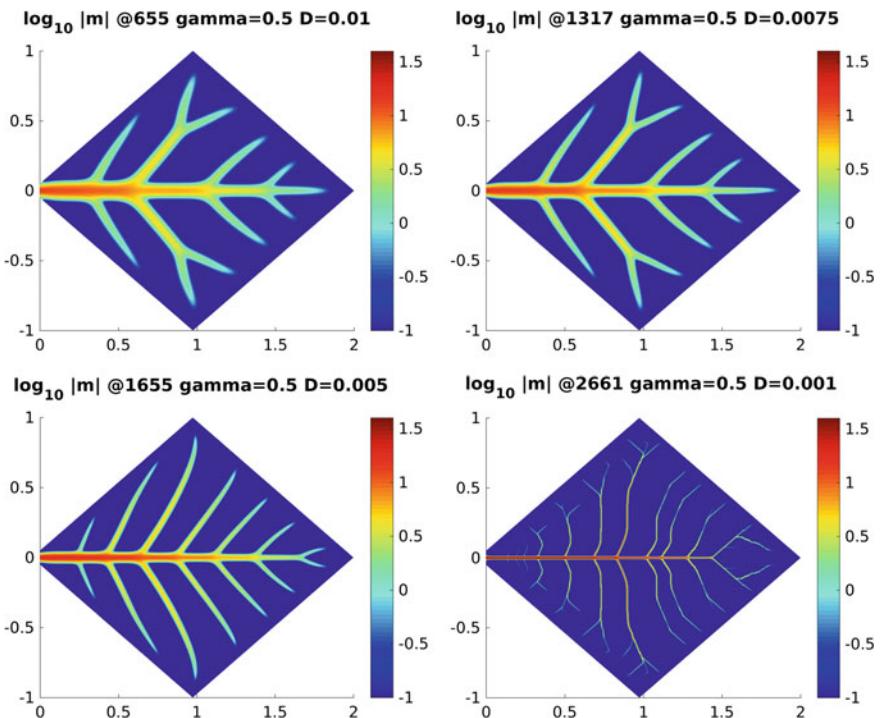


Fig. 8 Example 2: $\log_{10} |m|$ for $\gamma = 1/2$ and $D \in \{0.01, 0.0075, 0.005, 0.001\}$.

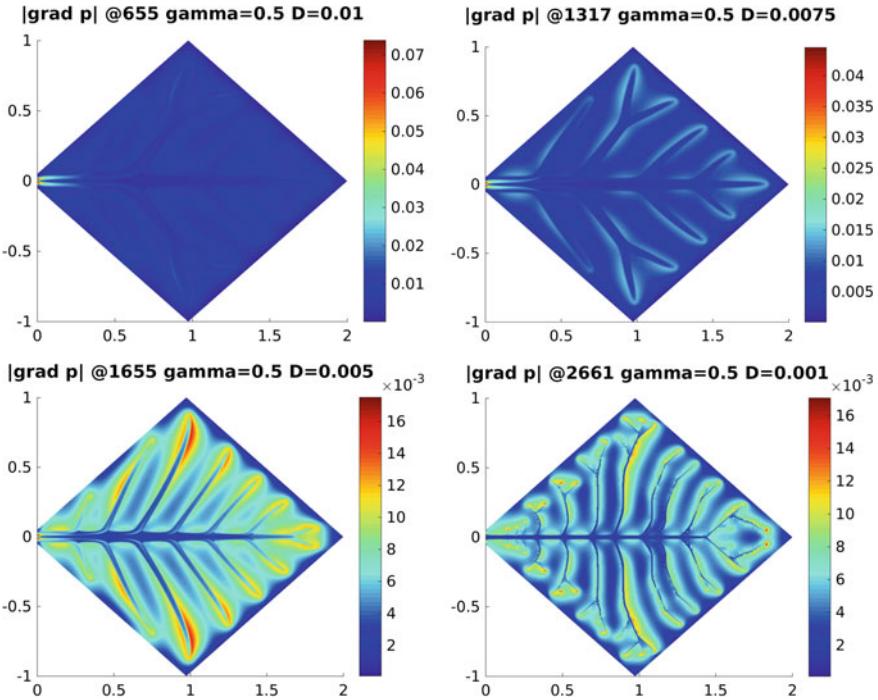


Fig. 9 Example 2: $|\nabla p|$ for $\gamma = 1/2$ and $D \in \{0.01, 0.0075, 0.005, 0.001\}$.

$\gamma > 0.5$, we set $\lambda_i^k = 0$ if $|\mathbf{m}_i^k| < 10^{-12}$. Furthermore, we solved (70) with preconditioned GMRES with a tolerance of 10^{-12} . We note that $\delta t_k \delta E^k \approx 10^{-12}$, here. Moreover, $\delta E^k / (\delta m^k)^2 < -1$ if $t^k \notin [2400, 2758]$, which shows sufficient decrease of the energy along the gradient flow. The situation is similar for $\gamma = 0.8$. Here, Algorithm 1 broke down at $t^k = 5899$ with corresponding energy $E^k = 0.411177$, and we started Algorithm 2 accordingly; see also Table 2.

In Figure 4, we show the resulting networks for $\gamma \in \{0.5, 0.75, 1\}$. For $\gamma \geq 0.75$, we observe that two straight lines connecting the source with the two sinks yield a local minimum of the energy functional. We remark that the situation is similar for $\gamma \in \{0.8, 0.9\}$, and therefore, we have not included the corresponding figures. We note that for $\gamma = 1$, $\|\nabla p\|_{L^2(\Omega)} \leq 1/c$ after the activation step of Algorithm 1 in the final iteration, which complies with the results of Section 3.3.3. The property $\|\nabla p\|_{L^2(\Omega)} \leq 1/c$ is not preserved in Algorithm 1 due to subsequent relaxation and diffusion steps. The corresponding values of $|\nabla p|$ are displayed in Figure 5. For $\gamma = 0.5$, we expect the iterates to develop into two straight lines as well. Let us refer to the results shown in [36] in this context. In view of [36], the final transportation network should have a branching point if $\gamma < 0.5$, i.e., a structure similar to the first picture in Figure 4. Figure 6 shows the direction of the flow u , i.e., the flow points

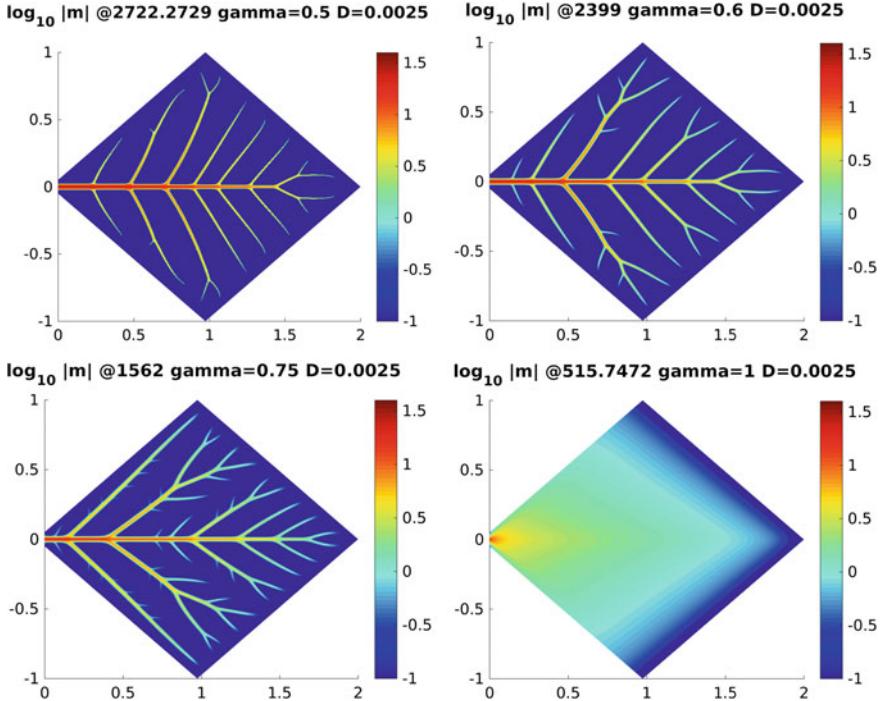


Fig. 10 Example 2: $\log_{10} |m|$ for $D = 0.0025$ and $\gamma \in \{0.5, 0.6, 0.75, 1\}$.

to the right where u_1 is positive, and it points to the top where u_2 is positive. In view of the choice of S and G , this complies with our expectation.

5.2 Example 2

We consider a diamond-shaped domain $\Omega \subset \mathbb{R}^2$ with one edge cut, see Figure 7, and a corresponding uniform triangulation with $M = 102\,905$ and $N = 204\,544$, i.e., $h \approx 3.165 \times 10^{-3}$. We define the initial datum m^I with components $m_2^I = 0$ and

$$m_1^I(x, y) = e^{-10^4 y^2 - 50x^2},$$

and define the boundary source and constant internal sinks as

$$G(x, y) = e^{-10^4(y^2+x^2)}, \quad S(x, y) = - \int_{\partial\Omega} G \, d\sigma / |\Omega|.$$

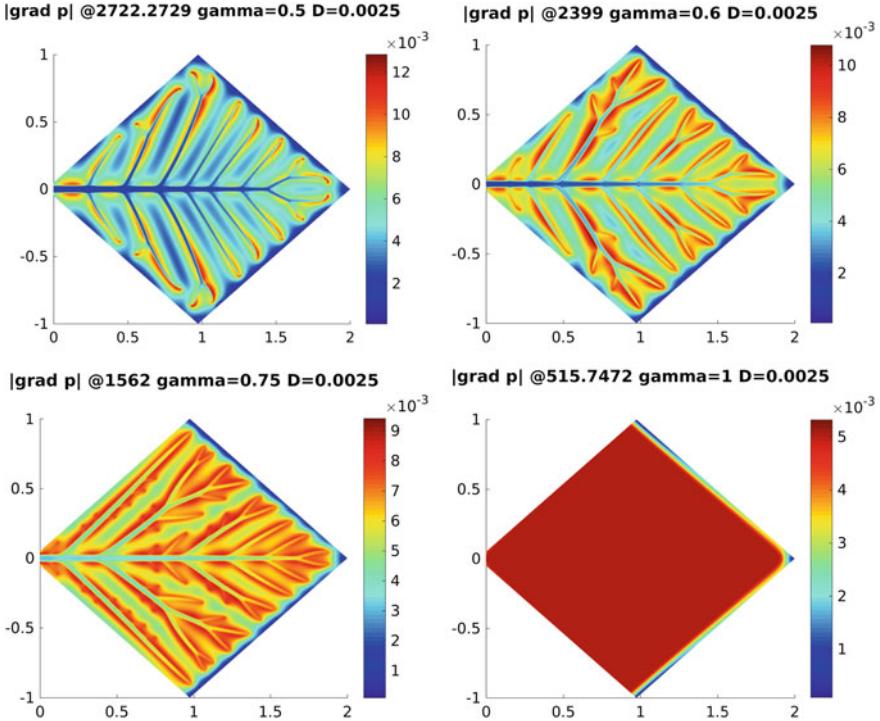


Fig. 11 Example 2: $|\nabla p|$ for $D = 0.0025$ and $\gamma \in \{0.5, 0.6, 0.75, 1\}$.

Hence, $\ell_{S,G}(1) = 0$. Since $\text{supp}(S) = \overline{\Omega}$, the whole domain Ω acts as a sink. The influence of the diffusion term is investigated next.

5.2.1 Varying D

The spreading of the network is strongly influenced by diffusion. While structures will be smoothed strongly for large diffusion, vanishing diffusion will rule out network formation. In Figure 8, we present different results of Algorithm 1 as described in Section 4.5 for $D \in \{0.01, 0.0075, 0.005, 0.001\}$ and $\gamma = 0.5$; for $D = 0.0025$ see Figure 10. As observed in Remark 10, Algorithm 1 may be interpreted as an approximate version of the forward–backward splitting algorithm, and, in the examples of this section, Algorithm 1 always decreased the energy, and the ratio $\delta E_h^k / (\delta m^k)^2 \leq -1$ indicates a sufficient decrease along the discrete gradient flow; cf. (82) and (31). We expect that for smaller diffusion, the resulting structures in the network get thinner, see Figure 8. Figure 9 shows the corresponding values of $|\nabla p^k|$. At the tip of the network $|\nabla p|$ is the largest, which shows that m grows along the gradient of the pressure. Furthermore, ∇p tends to zero, where network structures

Table 3 Example 2: Stationarity and sparsity measures for different values of D and $\gamma = 0.5$.

D	0.01	0.0075	0.005	0.0025	0.001
t^k	619	1317	1665	2722	2661
E^k	3.120	2.915	2.690	2.291	1.910
δm^k	2×10^{-3}	2×10^{-3}	2×10^{-3}	2×10^{-3}	4×10^{-4}
s_k	1.85	2.04	2.34	3.04	4.43
$c\ \nabla p^k\ _\infty$	14.76	8.91	3.50	2.57	3.42

Table 4 Example 2: Stationarity and sparsity measures for different values of γ and $D = 0.0025$.

γ	0.5	0.6	0.75	1
t^k	2722	2399	1562	515.7
E^k	2.291	2.540	3.021	3.651
δm^k	2×10^{-3}	2×10^{-3}	2×10^{-3}	7×10^{-4}
s_k	3.04	2.89	2.41	0.87
$c\ \nabla p\ _\infty$	2.57	2.16	1.88	1.06

have built up. In Table 3, we collected sparsity and stationarity measures. Although, our iterates have not converged to a stationary state yet, changes in energy and in the iterates are rather small; see also Figure 12 to get an impression of the changes in the network at this point of the evolution. We observe that the smaller D the larger the sparsity measure s_k is. We remark that for $D = 0.001$, $\delta t D^2 = 5 \times 10^{-7}$ is by more than one order smaller than $h^2 \approx 10^{-5}$. In view of the discussion in Section 4.3, the resolution of the diffusion process might already be a borderline case for $D = 0.001$.

5.2.2 Varying γ

The relaxation term plays an important role in the formation of networks as it models costs of a vessel. We present examples for $\gamma \in \{0.5, 0.6, 0.75, 1\}$ and $D = 0.0025$ in Figure 10. It becomes apparent that the closer γ is to 0.5 the sparser the network gets, i.e., the larger s_k is, see Table 4. This results in fewer vessels. In the other limiting case $\gamma = 1$, there is no specific spatial network structure. This is explained due to the fact that S is constant and $\text{supp}(S) = \overline{\Omega}$; for the case $\text{supp}(S) \neq \overline{\Omega}$, see Section 5.1. For completeness, we show in Figure 11 the results for $|\nabla p^k|$, which are similar to those in the previous section.

Figure 12 shows the evolution of $|m^k|$ for $\gamma = 0.6$ and $D = 0.0025$ for different times t^k . For small t^k , the conductivity m changes the most, and the structure is rather diffusive. The main branch, which is built up already in the first time steps, does not change much in the evolution compared to the other branches. For $t^k \geq 40$, the main structure has built up, and we observe a straightening of the secondary branches.

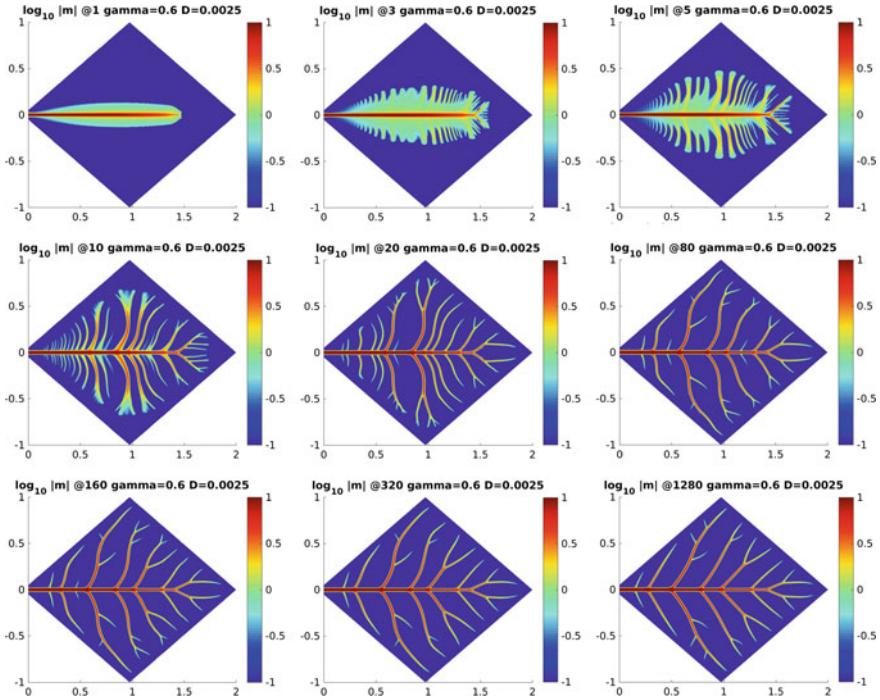


Fig. 12 Example 2: $\log_{10}(|m|)$ for $D = 0.0025$ and $\gamma = 0.6$ and different time steps.

For $t^k \geq 320$, it seems that mainly the tertiary branches change, i.e., vanish or get created; cf. also Figure 10.

Acknowledgements MB and MS acknowledge support by ERC via Grant EU FP 7 - ERC Consolidator Grant 615216 LifeInverse. MB acknowledges support by the German Science Foundation DFG via EXC 1003 Cells in Motion Cluster of Excellence, Münster, Germany. GA acknowledges the ERC-Starting Grant project High-Dimensional Sparse Optimal Control (HDSPCONTR).

References

1. G. Albi, M. Artina, M. Fornasier and P. Markowich: *Biological transportation networks: modeling and simulation*. Analysis and Applications. Vol. 14, Issue 01 (2016).
2. H. Attouch, J. Bolte and B. F. Svaiter: *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*. Math. Program., Ser. A (2013) 137:91–129.
3. J. Banavar, F. Colaiori, A. Flammini, A. Maritan and A. Rinaldo: *Topology of the Fittest Transportation Network*. Phys. Rev. Lett. 84, 20 (2000).
4. M. Barthélemy: *Spatial networks*. Physics Reports 499 (2011), pp. 1–101.
5. S. Bohn and M. Magnasco: *Structure, Scaling, and Phase Transition in the Optimal Transport Network*. Phys. Rev. Lett. 98, 088702 (2007).

6. M.A. Botchev: *A short guide to exponential Krylov subspace time integration for Maxwell's equations*. September 2012, Memorandum 1992, Department of Applied Mathematics, University of Twente, Enschede. ISSN 1874-4850.
7. S. C. Brenner, L. R. Scott, *The mathematical theory of finite element methods*, 3rd Edition, Vol. 15 of Texts in Applied Mathematics, Springer, New York, 2008.
8. H. Brezis: *New Questions Related to the Topological Degree*. In: P. Etingof, V. Retakh and I.M. Singer (eds.), *The Unity of Mathematics*, Progress in Mathematics 244, Birkhäuser Boston (2006), pp. 137–154.
9. L. Caffarelli, and N. Riviere, *The Lipschitz character of the stress tensor, when twisting an elastic plastic bar*. Arch. Rational Mech. Anal. 69 (1979), no. 1, pp. 3136.
10. P. Clément: *Approximation by finite element functions using local regularization*. Rev. Francaise Automat. Informat. Recherche Oérationnelle Sr. Rouge Anal. Numér., 9, R-2, 7784, 1975.
11. F. Corson: *Fluctuations and redundancy in optimal transport networks*. Phys. Rev. Lett. 104:048703 (2010).
12. N. Dengler and J. Kang: *Vascular patterning and leaf shape*. Current Opinion in Plant Biology 4 (2001), pp. 50–56.
13. G. Di Fazio: *L^p estimates for divergence form elliptic equations with discontinuous coefficients*. Boll. Un. Mat. Ital., (7) 10-A, pp. 409–420 (1996).
14. M. Durand: *Architecture of optimal transport networks*. Phys. Rev. E 73:016116 (2006).
15. J. Eckstein, D. P. Bertsekas: *On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators*. Math. Program., 55:293–318, 1992.
16. A. Eichmann, F. Le Noble, M. Autiero and P. Carmeliet: *Guidance of vascular and neural network formation*. Current Opinion in Neurobiology 15 (2005), pp. 108–115.?
17. M. Grant, S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming*, version 2.1, <http://cvxr.com/cvx>, 2014.
18. J.L. Gross and J. Yellen: *Handbook of Graph Theory*. CRC Press (2004).
19. D. Guidetti: *On elliptic systems in L^1* . Osaka J. Math. 30 (1993), pp. 397–429.
20. J. Haskovec, P. Markowich and B. Perthame: *Mathematical Analysis of a PDE System for Biological Network Formation*. Comm. PDE 40:5, pp. 918–956, 2015.
21. J. Haskovec, S. Hittmeir, P. Markowich and A. Mielke: *On energy-reaction-diffusion systems*. Preprint, 2016.
22. J. Haskovec, P. Markowich, B. Perthame and M. Schlottbom: Notes on a PDE system for biological network formation. Nonlinear Analysis 138 (2016), pp. 127–155.
23. https://en.wikipedia.org/wiki/Intensive_and_extensive_properties.
24. D. Hu: *Optimization, Adaptation, and Initialization of Biological Transport Networks*. Notes from lecture (2013).
25. D. Hu and D. Cai: *Adaptation and Optimization of Biological Transport Networks*. Phys. Rev. Lett. 111 (2013), 138701.
26. D. Hu, D. Cai and W. Rangan: *Blood Vessel Adaptation with Fluctuations in Capillary Flow Distribution*. PLoS ONE 7:e45444 (2012).
27. F. John and L. Nirenberg: *On functions of bounded mean oscillation*. Comm. Pure and appl. 14 (1961), pp. 415–426.
28. A. Kamiya, R. Bukhari and T. Togawa: *Adaptive regulation of wall shear stress optimizing vascular tree function*. Bulletin of Mathematical Biology 46 (1984), pp. 127–137.
29. E. Katifori, G. Szollosi and M. Magnasco: *Damage and fluctuations induce loops in optimal transport networks*. Phys. Rev. Lett. 104:048704 (2010).
30. M. Laguna, S. Bohn and E. Jagla: *The Role of Elastic Stresses on Leaf Venation Morphogenesis*. PLOS Comput. Biol. 4:e1000055 (2008).
31. R. Malinowski: *Understanding of Leaf Developmentthe Science of Complexity*. Plants 2 (2013), pp. 396–415.
32. N. Meyers: *An L^p -estimate for the gradient of solutions of second order elliptic divergence equations*. Annali della Scuola Normale Superiore di Pisa, Classe di Scienze 3^e série, tome 17, no. 3 (1963), pp. 189–206.

33. O. Michel and J. Biondi: *Morphogenesis of neural networks*. Neural Processing Letters, Vol. 2, No. 1 (1995), pp. 9–12.
34. C. Murray: *The physiological principle of minimum work. I. The vascular system and the cost of blood volume*. Proc. Natl. Acad. Sci. USA. 12, 207 (1926).
35. T. Nelson and N. Dengler: *Leaf Vascular Pattern Formation*. Plant Cell 9:1121 (1997).
36. E. Oudet and F. Santambrogio: *A Modica-Mortola Approximation for Branched Transport and Applications*. Archive for Rational Mechanics and Analysis. Vol. 201, Issue 1, pp. 115–142 (2011).
37. N. Parikh and S. Boyd: *Proximal Algorithms*. Foundations and Trends in Optimization. Vol. 1, No. 3 (2013) 123231.
38. U. Pohl, J. Holtz, R. Busse and E. Bassenge: *Crucial role of endothelium in the vasodilator response to increased flow in vivo*. Hypertension 8:37–44 (1986).
39. M. Safdari: *The free boundary of variational inequalities with gradient constraints*. arXiv:1501.05337.
40. M. Safdari: *The regularity of some vector-valued variational inequalities with gradient constraints*. arXiv:1501.05339.
41. D. Sarason: *Functions of vanishing mean oscillation*. Trans. AMS 207 (1975), pp. 391–405.
42. D. Sedmera: *Function and form in the developing cardiovascular system*. Cardiovascular Research 91 (2011), pp. 252–259.
43. J. Simon: *Compact sets in the space $L^p(0, T; B)$* . Ann. Mat. Pure Appl. IV (146), 1987, pp. 65–96.
44. A. Tero, S. Takagi, T. Saigusa, K. Ito, D. Bebber, M. Fricker, K. Yumiki, R. Kobayashi and T. Nakagaki: *Rules for Biologically Inspired Adaptive Network Design*. Science 22:327 (2010), pp. 439–442.
45. C. K. Weichert: *Cardiovascular System*. AccessScience. McGraw-Hill Education (2014).
46. S. Whitaker: *Flow in porous media I: A theoretical derivation of Darcys law*. Transport in porous media, 1 (1986), pp. 3–25.

Recent Advances in Opinion Modeling: Control and Social Influence

Giacomo Albi, Lorenzo Pareschi, Giuseppe Toscani
and Mattia Zanella

Abstract We survey some recent developments on the mathematical modeling of opinion dynamics. After an introduction on opinion modeling through interacting multi-agent systems described by partial differential equations of kinetic type, we focus our attention on two major advancements: optimal control of opinion formation and influence of additional social aspects, like conviction and number of connections in social networks, which modify the agents' role in the opinion exchange process.

1 Preliminaries

We introduce some of the essential literature on the opinion formation, by emphasizing the role of the kinetic description. New problems recently treated in the scientific community are outlined. Then, the mathematical description of the core ideas of kinetic models for opinion formation are presented in details.

G. Albi
Technische Universität München, Boltzmannstraße 3,
Garching (München), Germany
e-mail: giacomo.albi@ma.tum.de

L. Pareschi · M. Zanella
University of Ferrara, via Machiavelli 35, Ferrara, Italy
e-mail: lorenzo.pareschi@unife.it

M. Zanella
e-mail: mattia.zanella@unife.it

G. Toscani (✉)
University of Pavia, via Ferrata 1, Pavia, Italy
e-mail: giuseppe.toscani@unipv.it

1.1 Introduction

The statistical physics approach to social phenomena is currently attracting much interest, as indicated by the huge and rapidly increasing number of papers and monographies based on it [16, 37, 86, 89]. In this rapidly increasing field of research, because of its pervasiveness in everyday life, the process of opinion formation is nowadays one of the most studied application of mathematics to social dynamics [18, 25, 28, 58, 72, 78, 79, 98].

Along this survey, we focus on some recent advances in opinion formation modeling, which aims in coupling the process of opinion exchange with other aspects, which are closely related to the process itself, and takes into account the dependence on new variables which are usually neglected, mainly in reason of the mathematical difficulties that the introduction of further dimensions add to the models.

These new aspects are deeply connected and range from opinion leadership and opinion control, to the role of conviction and the interplay between complex networks and the spreading of opinions. In fact, leaders are recognized to be important since they can exercise control over public opinion. It is a concept that goes back to Lazarsfeld et al. [75]. In the course of their study of the presidential elections in the USA in 1940, it was found interpersonal communication to be much more influential than direct media effects. In [75] a theory of a two-step flow of communication was formulated, where so-called opinion leaders who are active media users select, interpret, modify, facilitate and finally transmit information from the media to less active parts of the population. It is clear that various principal questions arise, mainly linked to this two-step flow of communication. The first one is related to the ability to effectively exercise a control on opinion and to the impact of modern communication systems, like social networks, to the dynamics of opinions. The second is related to the fact that the less active part of the population is in general adapting to leaders opinion only partially. Indeed, conviction plays a major role in this process, by acting as a measurable resistance to the change of opinion.

These enhancements will be modeled using the toolbox of classical kinetic theory [89]. Within this choice, one will be able to present an almost uniform picture of opinion dynamics, starting from few simple rules. The kinetic model of reference was introduced by one of the authors in 2006 [98], and was subsequently generalized in many ways (see [86, 89] for recent surveys). The building block of kinetic models are pairwise interactions. In classical opinion formation, interactions among agents are usually described in terms of few relevant concepts, represented respectively by compromise and self-thinking. Once fixed in binary interactions, the microscopic rules are responsible of the formation of coherent structures.

The remarkably simple compromise process describes mathematically the way in which pairs of agents reach a fair compromise after exchanging opinions. The rule of compromise has been intensively studied [18–20, 51, 72, 104]. The second one is the self-thinking process, which allows individual agents to change their opinions in an unpredictable way. It is usually mathematically described in terms of some random variable [18, 98]. The resulting kinetic models are sufficiently general to

take into account a large variety of human behaviors, and to reproduce in many cases explicit steady profiles, from which one can easily elaborate information on the opinion behavior [11, 27–30]. For the sake of completeness, let us mention that many other models with analogous properties have been introduced and studied so far [13, 17, 23, 25, 43, 54, 57, 64–66, 74, 91, 94, 97, 99, 100, 103].

Kinetic models have been also the basis for suitable generalizations, in which the presence of leaders and their effect of the opinion dynamics has been taken into account [4, 32, 41, 42, 45, 58, 59]. Also, the possibility to establish an effective control on opinion, both through an external media or through the leaders' action, has attracted the interest of the research community [6, 7, 21, 22]. The methods here are strongly connected to analogous studies in crowd dynamics and flocking phenomena [5, 8, 26, 36, 46, 49, 60, 67].

Further, the effect of conviction in the formation of opinion started to be studied. While in general conviction is assumed to appear as a static parameter in the opinion dynamics [47, 48, 84, 104], in [31] conviction has been assumed to follow a proper evolution in the society on the basis of interactions with an external background. Recently, a similar approach has been used to model the effect of competence and the so-called equality bias phenomena [91]. This point of view was previously applied to the study of the formation of knowledge in [90], as a starting point to investigate its role in wealth distribution. Indeed, many of the aforementioned models share a common point of view with the statistical approach to distribution of wealth [38, 44, 86, 90].

More recently, in reason of their increasing relevance in modern societies, the statistical mechanics of opinion formation started to be applied to extract information from complex social networks [1, 2, 12, 15, 25, 69, 70, 87, 92, 96, 102]. In these models the number of connections of the agents play a major role in characterizing the dynamics [9, 10, 50, 56, 68]. In [9] the model links the graph evolution modeled by a discrete connection distribution dynamics with the spreading of opinion along the network.

Before starting our survey, it is essential to outline the peculiar aspects of the microscopic details of the binary interactions which express the microscopic change of opinion. Indeed, these interactions differ in many aspects from the usual binary interactions considered in classical kinetic theory of rarefied gases. The first difference is that opinion is usually identified with a continuum variable which can take values in a bounded interval. Second, the post-interaction opinions are not a linear transformation of the pre-interaction ones. Indeed, it is realistic to assume that people with a neutral opinion is more willing to change it, while the opposite phenomenon happens with people which have extremal opinions.

Once the details of the pairwise interactions are fixed, the explicit form of the bilinear kinetic equation of Boltzmann type follows [89]. One of the main consequences of the kinetic description is that it constitute a powerful starting point to obtain, in view of standard asymptotic techniques, continuous mean-field models with a reduced complexity, which maintain most of the physical properties of the underlying Boltzmann equation. The main idea, is to consider important only interactions which are *grazing*, namely interactions in which the opinion variable does not

change in a sensible way, while at the same time the frequency of the interactions is assumed to increase. This asymptotic limit (hereafter called quasi-invariant opinion limit) leads to partial differential equations of Fokker-Planck type for the distribution of opinion among individuals, that in many cases allow for an analytic study.

The rest of the survey is organized as follows. In the remaining part of Section 1 we describe the basic kinetic model for opinion formation. It represents the building block for the binary opinion dynamics which is used in the subsequent Sections. Next in Section 2 we deal with control problems for opinion dynamics. First by considering an external action which forces the agents towards a desired state and subsequently by introducing a leaders' population which acts accordingly to a prescribed optimal strategy. Here we start from the optimal control problem for the corresponding microscopic dynamics and approximate it through a finite time horizon or model predictive control technique. This permits to embed the feedback control directly into the limiting kinetic equations. Section 3 is then devoted to the modeling through multivariate distribution functions where the agents' opinion is coupled with additional variables. Specifically we consider the case where conviction is also an evolving quantity playing a role in the dynamics and the case where agents interact over an evolving social network accordingly to their number of connections. Some final remarks are contained in the last Section and details on numerical methods are given in a separate Appendix.

1.2 Kinetic Modeling

On the basis of statistical mechanics, to construct a model for opinion formation the fundamental assumption is that agents are indistinguishable [89]. An agent's *state* at any instant of time $t \geq 0$ is completely characterized by his opinion $w \in [-1, 1]$, where -1 and 1 denote two (extreme) opposite opinions.

The unknown is the density (or distribution function) $f = f(w, t)$, where $w \in I = [-1, 1]$ and the time $t \geq 0$, whose time evolution is described, as shown later, by a kinetic equation of Boltzmann type.

The precise meaning of the density f is the following. Given the population to study, if the opinions are defined on a sub-domain $D \subset \mathfrak{I}$, the integral

$$\int_D f(w, t) dw$$

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

$$\int_I f(w, t) dw = 1.$$

As always happens when dealing with a kinetic problem in which the variable belongs to a bounded domain, this choice introduces supplementary mathematical difficulties in the correct definition of binary interactions. In fact, it is essential to consider only interactions that do not produce opinions outside the allowed interval, which corresponds to imposing that the extreme opinions cannot be crossed. This crucial limitation emphasizes the difference between the present *social* interactions, where not all outcomes are permitted, and the classical interactions between molecules, or, more generally, the wealth trades (cf. [89], Chapter 5), where the only limitation for trades was to insure that the post-collision wealths had to be non-negative.

In order to build a realistic model, this severe limitation has to be coupled with a reasonable physical interpretation of the process of opinion forming. In other words, the impossibility of crossing the boundaries has to be a by-product of good modeling of binary interactions.

From a microscopic viewpoint, the binary interactions in [98] were described by the rules

$$\begin{aligned} w' &= w - \eta P(w, w_*)(w - w_*) + \xi D(w), \\ w'_* &= w_* - \eta P(w_*, w)(w_* - w) + \xi_* D(w_*). \end{aligned} \tag{1}$$

In (1), the pair (w, w_*) , with $w, w_* \in I$, denotes the opinions of two arbitrary individuals before the interaction, and (w', w'_*) their opinions after exchanging information between each other and with the exterior. The coefficient $\eta \in (0, 1/2)$ is a given constant, while ξ and ξ_* are random variables with the same distribution, with zero mean and variance ς^2 , taking values on a set $\mathcal{B} \subseteq \mathbb{R}$. The constant η and the variance ς^2 measure respectively the compromise propensity and the degree of spreading of opinion due to diffusion, which describes possible changes of opinion due to personal access to information (self-thinking). Finally, the functions $P(\cdot, \cdot)$ and $D(\cdot)$ take into account the local relevance of compromise and diffusion for given opinions.

Let us describe in detail the interaction on the right-hand side of (1). The first part is related to the compromise propensity of the agents, and the last contains the diffusion effects due to individual deviations from the average behavior. The presence of both the functions $P(\cdot, \cdot)$ and $D(\cdot)$ is linked to the hypothesis that openness to change of opinion is linked to the opinion itself, and decreases as one gets closer to extremal opinions. This corresponds to the natural idea that extreme opinions are more difficult to change. Various realizations of these functions can be found in [89, 98]. In all cases, however, we assume that both $P(w, w_*)$ and $D(w)$ are non-increasing with respect to $|w|$, and in addition $0 \leq P(w, w_*) \leq 1$, $0 \leq D(w) \leq 1$. Typical examples are given by $P(w, w_*) = 1 - |w|$ and $D(w) = \sqrt{1 - w^2}$.

In the absence of the diffusion contribution ($\xi, \xi_* \equiv 0$), (1) implies

$$\begin{aligned} w' + w'_* &= w + w_* + \eta(w - w_*)(P(w, w_*) - P(w_*, w)), \\ w' - w'_* &= (1 - \eta(P(w, w_*) + P(w_*, w))) (w - w_*). \end{aligned} \tag{2}$$

Thus, unless the function $P(\cdot, \cdot)$ is assumed constant, $P = 1$, the *mean opinion* is not conserved and it can increase or decrease depending on the opinions before the interaction. If $P(\cdot, \cdot)$ is assumed constant, the conservation law is reminiscent of analogous conservations which take place in kinetic theory. In such a situation, thanks to the upper bound on the coefficient η , equations (1) correspond to a granular-gas-like interaction [89] where the stationary state is a Dirac delta centered on the average opinion. This behavior is a consequence of the fact that, in a single interaction, the compromise propensity implies that the difference of opinion is diminishing, with $|w' - w'_*| = (1 - 2\eta)|w - w_*|$. Thus, all agents in the society will end up with exactly the same opinion.

We remark, moreover, that, in the absence of diffusion, the lateral bounds are not violated, since

$$\begin{aligned} w' &= (1 - \eta P(w, w_*))w + \eta P(w, w_*)w_*, \\ w'_* &= (1 - \eta P(w_*, w))w_* + \eta P(w_*, w)w, \end{aligned} \tag{3}$$

imply

$$\max \{|w'|, |w'_*|\} \leq \max \{|w|, |w_*|\}.$$

Let $f(w, t)$ denote the distribution of opinion $w \in I$ at time $t \geq 0$. The time evolution of f is recovered as a balance between bilinear gain and loss of opinion terms, described in weak form by the integro-differential equation of Boltzmann type

$$\begin{aligned} \frac{d}{dt} \int_I \varphi(w) f(w, t) dv &= (Q(f, f), \varphi) = \\ \lambda \left\langle \int_{I^2} (\varphi(w') + \varphi(w'_*) - \varphi(w) - \varphi(w_*)) f(w) f(w_*) dw dw_* \right\rangle, \end{aligned} \tag{4}$$

where (w', w'_*) are the post-interaction opinions generated by the pair (w, w_*) in (1), λ represents a constant rate of interaction and the brackets $\langle \cdot \rangle$ denote the expectation with respect to the random variables ξ and ξ_* .

Equation (4) is consistent with the fact that a suitable choice of the function $D(\cdot)$ in (1) coupled with a small support \mathcal{B} of the random variables implies that both $|w'| \leq 1$ and $|w'_*| \leq 1$. We do not insist here on further details on the evolution properties of the solution, by leaving them to the next Sections, where these properties are studied for the particular problems.

2 Optimal Control of Consensus

Different to the classical approach where individuals are assumed to freely interact and exchange opinions 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 kinetic models including those controls. This can be used to study the influence on the system dynamics to enforce emergence of non spontaneous desired asymptotic states.

Two relevant situations will be explored, first a distributed control, which models the action of an external force acting as a *policy maker*, like the effects of the media [6], next an indirect internal control, where we assume that the control corresponds to the strategies of *opinion leaders*, aiming to influence the consensus of the whole population [7]. In order to characterize the kinetic structure of the optimal control of consensus dynamics, we will start to derive it as a feedback control from a general optimal control problem for the corresponding microscopic model, and thereafter we will connect it to the binary dynamics.

2.1 Control by an External Action

We consider the microscopic evolution of the opinions of N agents, where each agent's opinion $w_i \in I$, $I = [-1, 1]$, $i = 1, \dots, N$ evolves according the following first order dynamical system

$$\dot{w}_i = \frac{1}{N} \sum_{j=1}^N P(w_i, w_j)(w_j - w_i) + u, \quad w_i(0) = w_{0,i}, \quad (5)$$

where $P(\cdot, \cdot)$ has again the role of the compromise function defined in (1). The control $u = u(t)$ models the action of an external agent, e.g. a *policy maker* or social media. We assume that it is the solution of the following optimal control problem

$$u = \arg \min_{u \in \mathcal{U}} J(u) := \frac{1}{2} \int_0^T \frac{1}{N} \sum_{j=1}^N ((w_j - w_d)^2 + \kappa u^2) ds, \quad u(t) \in [u_L, u_R], \quad (6)$$

with \mathcal{U} the space of the admissible controls. In formulation (6) a quadratic cost functionals with a penalization parameter $\kappa > 0$ is considered, and the value w_d represents the desired opinion state. We refer to [3, 6, 36, 60] for further discussion on the analytical and numerical studies on this class of problems. The additional constraints on the pointwise values of $u(t)$ given by u_L and u_R , are necessary in order to preserve the bounds of $w_i \in I$ (see [35, 82]).

2.1.1 Model Predictive Control of the Microscopic Dynamics

In general, for large values of N , standard methods for the solution of problems of type (5)–(6) over the full time interval $[0, T]$ stumble upon prohibitive computational costs due to the nonlinear constraints.

In what follows we sketch an approximation method for the solution of (5)–(6), based on *model predictive control* (MPC), which furnishes a suboptimal control by an iterative solution over a sequence of finite time steps, but, nonetheless, it allows an explicit representation of the control strategy [6, 35, 40, 73, 80, 81, 95]. The link between MPC on the level of agents and the MPC on the level of kinetic and fluid-dynamic equations has been subject to recent investigations in [6, 7, 71], and also the relation between MPC and mean-field games has been studied in [34, 52, 55, 77].

Let us consider the time sequence $0 = t_0 < t_1 < \dots < t_M = T$, a discretization of the time interval $[0, T]$, where $\Delta t = t_n - t_{n-1}$, for all $n = 1, \dots, M$ and $t_M = M \Delta t$. Then we assume the control to be constant on every interval $[t_n, t_{n+1}]$, and defined as a piecewise function, as follows

$$\bar{u}(t) = \sum_{n=0}^{M-1} \bar{u}^n \chi_{[t_n, t_{n+1}]}(t), \quad (7)$$

where $\chi(\cdot)$ is the characteristic function of the interval $[t_n, t_{n+1}]$. We consider a full discretization of the optimal control problem (5)–(6), through a forward Euler scheme, and we solve on every time frame $[t_n, t_n + \Delta t]$, the reduced optimal control problem

$$\min_{\bar{u} \in \bar{\mathcal{U}}} J_{\Delta t}(\bar{u}) := \frac{1}{2N} \sum_{j=1}^N (w_j^{n+1} - w_d)^2 + \frac{\kappa}{2} \int_{t_n}^{t_{n+1}} \bar{u}^2 dt, \quad (8)$$

subject to

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{N} \sum_{j=1}^N P(w_i^n, w_j^n)(w_j^n - w_i^n) + \Delta t \bar{u}^n, \quad w_i^n = w_i(t_n), \quad (9)$$

for all $i = 1, \dots, N$, and \bar{u} in the space of the admissible controls $\bar{\mathcal{U}} \subset \mathcal{U}$. Note that since the control \bar{u} is a constant value over the time interval $[t_n, t_n + \Delta t]$, and w^{n+1} depends linearly on \bar{u}^n through (9), the discrete optimal control problem (8) reduces to

$$J_{\Delta t}(\bar{u}^n) = \frac{1}{2N} \sum_{j=1}^N (w_j^{n+1}(\bar{u}^n) - w_d)^2 + \Delta t \frac{\kappa}{2} (\bar{u}^n)^2. \quad (10)$$

Thus, in order to find the minimizer of (8), it is sufficient to compute the derivative of (10) with returns us the optimal value expressed as follows

$$U^n = -\frac{1}{\kappa + \Delta t} \left(\frac{1}{N} \sum_{j=1}^N (w_j^n - w_d) + \frac{\Delta t}{N^2} \sum_{j,k} P(w_j^n, w_k^n)(w_k^n - w_j^n) \right). \quad (11)$$

Expression (11) furnishes a feedback control for the full discretized problem, which can be plugged as an *instantaneous control* into (9), obtaining the following constrained system

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{N} \sum_{j=1}^N P(w_i^n, w_j^n)(w_j^n - w_i^n) + \Delta t U^n, \quad w_i^n = w_i(t_n). \quad (12)$$

A more general derivation can be obtained through a discrete Lagrangian approach for the optimal control problem (8)–(9), see [6].

Remark 1 We remark that the scheme (12) furnishes a suboptimal solution w.r.t. the original control problem. In particular if $P(\cdot, \cdot)$ is symmetric, only the average of the system is controlled. Let us set $w_d = 0$, and $m^n = \sum_{i=1}^N w_i^n / N$. Summing on $i = 1, \dots, N$ equation (12) we have

$$m^{n+1} = m^n - \frac{\Delta t}{\kappa + \Delta t} m^n = \left(1 - \frac{\Delta t}{\kappa + \Delta t} \right)^n m^0, \quad (13)$$

which implies $m^\infty = 0$. Thus, while the feedback control is able to control the mean of the system, it does not assure the global consensus. We will see in the next Section how the introduction of a binary control depending on the pairs permits to recover the global consensus.

2.1.2 Binary Boltzmann Control

Following Section 1.2, we consider now a kinetic model for the evolution of the density $f = f(w, t)$ of agents with opinion $w \in I = [-1, 1]$ at time $t \geq 0$, such that the total mass is normalized to one. The evolution can be derived by considering the change in time of $f(w, t)$ depending on the interactions among the individuals of the binary type (1). In order to derive such Boltzmann description we follow the approach in [5, 61]. We consider the model predictive control system (12) in the simplified case of only two interacting agents, numbered i and j . Their opinions are modified according to

$$\begin{aligned} w_i^{n+1} &= w_i^n + \frac{\Delta t}{2} P(w_i^n, w_j^n)(w_j^n - w_i^n) + \frac{\Delta t}{2} U(w_i^n, w_j^n), \\ w_j^{n+1} &= w_j^n + \frac{\Delta t}{2} P(w_i^n, w_j^n)(w_i^n - w_j^n) + \frac{\Delta t}{2} U(w_j^n, w_i^n), \end{aligned} \quad (14)$$

where the feedback control term $U(w_i^n, w_j^n)$ is derived from (11) and yields

$$\begin{aligned} \frac{\Delta t}{2} U(w_i^n, w_j^n) = & -\frac{1}{2} \frac{\Delta t}{\kappa + \Delta t} ((w_j^n - w_d) + (w_i^n - w_d)) \\ & - \frac{1}{4} \frac{\Delta t^2}{\kappa + \Delta t} (P_{ij}^n - P_{ji}^n) (w_j^n - w_i^n), \end{aligned} \quad (15)$$

having defined $P_{ij}^n = P(w_i^n, w_j^n)$. This formulation can be written as a binary Boltzmann dynamics

$$\begin{aligned} w' &= w + \eta P(w, w_*) (w - w_*) + \eta U(w, w_*) + \xi D(w), \\ w'_* &= w_* + \eta P(w_*, w) (w_* - w) + \eta U(w_*, w) + \xi_* D(w_*). \end{aligned} \quad (16)$$

All quantities in (16) are defined as in (1). The control $U(\cdot, \cdot)$, which is not present in (1), acts as a forcing term to steer consensus, or, in other words, it models the action of promoting the emergence of a desired status.

Thus we can associate the binary dynamics in (14) to the original dynamics in (16). Choosing $\eta = \Delta t/2$, the control term for the arbitrary pair (w, w_*) reads

$$\eta U(w, w_*) = \frac{2\eta}{\kappa + 2\eta} (K(w, w_*) + \eta H(w, w_*)), \quad (17)$$

where

$$K(w, w_*) = \frac{1}{2} ((w_d - w) + (w_d - w_*)), \quad (18)$$

$$H(w, w_*) = \frac{1}{2} (P(w, w_*) - P(w_*, w)) (w - w_*). \quad (19)$$

Note that $K(w, w_*)$ and $H(w, w_*)$ are both symmetric, which follows directly by (5)–(6), since u is the same for every agent. Embedding the control dynamics into (16) we obtain the following binary constrained interaction

$$\begin{aligned} w' &= w + \eta P(w, w_*) (w_* - w) + \beta(K(w, w_*) + \eta H(w, w_*)) + \xi D(w), \\ w'_* &= w_* + \eta P(w_*, w) (w - w_*) + \beta(K(w, w_*) + \eta H(w, w_*)) + \xi_* D(w_*), \end{aligned} \quad (20)$$

with β defined as follows

$$\beta := \frac{2\eta}{\kappa + 2\eta}. \quad (21)$$

In the absence of diffusion, from (20) it follows that

$$\begin{aligned} w' + w'_* &= w + w_* + \eta(P(w, w_*) - P(w_*, w))(w_* - w) \\ &\quad + 2\beta(K(w, w_*) + \eta H(w, w_*)) \\ &= w + w_* - 2\eta H(w, w_*) + 2\beta(K(w, w_*) + \eta H(w, w_*)) \\ &= 2w_d - (1 - \beta)(\kappa + 2\eta)U(w, w_*) = 2w_d - \kappa U(w, w_*), \end{aligned} \quad (22)$$

thus in general the mean opinion is not conserved. Observe that the computation of the relative distance between opinions $|w' - w'_*|$ is equivalent to (2), since the subtraction cancels the control terms out, giving the inequality

$$|w' - w'_*| = (1 - \eta(P(w, w_*) + P(w_*, w)))|w - w_*| \leq (1 - 2\eta)|w - w_*|, \quad (23)$$

which tells that the relative distance in opinion between two agents diminishes after each interaction [98]. In presence of noise terms, we should assure that the binary dynamics (20) preserves the boundary, i.e. $w', w'_* \in I$. An important role in this is played by functions $P(\cdot, \cdot)$, $D(\cdot)$, as stated by the following proposition.

Proposition 1 *Let assume that there exist $p > 0$ and m_C such that $p \leq P(w, w_*) \leq 1$ and $m_C = \min\{(1 - w)/D(w), D(w) \neq 0\}$. Then, provided*

$$\beta \leq np, \quad \Theta \in \left(-m_C \left(\eta - \frac{\beta}{2}\right), m_C \left(\eta - \frac{\beta}{2}\right)\right), \quad (24)$$

are satisfied, the binary interaction (20) preserves the bounds, i.e. the post-interaction opinions w', w'_ are contained in $I = [-1, 1]$.*

Proof We refer to [6, 98] for a detailed proof.

Remark 2 Observe that, from the modeling viewpoint, noise is seen as an external term which can not be affected by the control dynamics. A different strategy is to account the action of the noise at the level of the microscopic dynamics (5) and proceed with the optimization. This will lead to a different binary interaction with respect to (20), where the control influences also the action of the noise.

2.1.3 Main Properties of the Boltzmann Description

In general the time evolution of the density $f(w, t)$ is found by resorting to a Boltzmann equation of type (4), where the collisions are now given by (20). In weak form we have

$$(Q(f, f), \varphi) = \frac{\lambda}{2} \left\langle \int_{I^2} (\varphi(w') + \varphi(w_*) - \varphi(w) - \varphi(v)) f(w) f(w_*) dw dw_* \right\rangle, \quad (25)$$

where we omitted the time dependence for simplicity. Therefore the total opinion, obtained taking $\varphi(w) = 1$, is preserved in time. This is the only conserved quantity of the process. Choosing $\varphi(w) = w$, we obtain the evolution of the average opinion, thus we have

$$\frac{d}{dt} \int_I w f(w, t) dw = \frac{\lambda}{2} \left\langle \int_{I^2} (w' + w'_* - w - w_*) f(w, t) f(v, t) dw dw_* \right\rangle. \quad (26)$$

Indicating the average opinion as $m(t) = \int_I w f(w, t) dw$, using (22) we get

$$\begin{aligned} \frac{dm(t)}{dt} &= \lambda \beta (w_d - m(t)) \\ &+ \lambda \eta (1 - \beta) \int_{I^2} (P(w, w_*) - P(w_*, w)) w_* f(w_*) f(w) dw dw_*. \end{aligned} \quad (27)$$

Since $0 \leq P(w, w_*) \leq 1$, $|P(w, w_*) - P(w_*, w)| \leq 1$, we can bound the derivative from below and above

$$\lambda \beta w_d - \lambda(\beta + \eta(1 - \beta))m(t) \leq \frac{d}{dt} m(t) \leq \lambda \beta w_d - \lambda(\beta - \eta(1 - \beta))m(t).$$

Note that in the limit $t \rightarrow \infty$, the average $m(t)$ converges to the desired state w_d , if $\beta - \eta(1 - \beta) > 0$. This implies the following restriction $\kappa < 2$. A similar analysis can be performed for the second moment $\varphi(w) = w^2$, showing the decay of the energy for particular choices of the interaction potential (cf. [6, 7, 98] for further details on the proprieties of the moment of (4)).

Remark 3 In the symmetric case, $P(v, w) = P(w, v)$, equation (27) is solved explicitly as

$$m(t) = (1 - e^{-\lambda \beta t}) w_d + m(0) e^{-\lambda \beta t} \quad (28)$$

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

2.1.4 The Quasi-invariant Opinion Limit

We will now introduce some asymptotic limit of the kinetic equation. The main idea is to scale interaction frequency and strength, λ and η respectively, diffusion ς^2 at the same time, in order to maintain at any level of scaling the memory of the microscopic interactions (20). This approach is referred to as *quasi-invariant opinion limit* [63, 98, 101]. Given $\varepsilon > 0$, we consider the following scaling

$$\eta = \varepsilon, \quad \lambda = \frac{1}{\varepsilon}, \quad \varsigma = \sqrt{\varepsilon} \sigma, \quad \beta = \frac{2\varepsilon}{\kappa + 2\varepsilon}. \quad (29)$$

The ratio $\varsigma^2/\eta = \sigma$ is of paramount importance in order to show in the limit the contribution of both the compromise propensity η and the diffusion ς^2 . Other scalings lead to diffusion dominated ($\varsigma^2/\eta \rightarrow \infty$) or compromise dominated ($\varsigma^2/\eta \rightarrow 0$) equations. In the sequel we show through formal computations how this approach leads to a Fokker–Planck equation type [93]. We refer to [98] for details and rigorous derivation.

After scaling, equation (25) reads

$$\frac{d}{dt} \int_I \varphi(w) f(w, t) dw = \frac{1}{\varepsilon} \left\langle \int_{I^2} (\varphi(w') - \varphi(w)) f(w, t) f(v, t) dw dv \right\rangle, \quad (30)$$

while the scaled binary interaction dynamics (20) is given by

$$w' - w = \varepsilon P(w, w_*) (w_* - w) + \frac{2\varepsilon}{\kappa + 2\varepsilon} K(w, w_*) + \xi D(w) + O(\varepsilon^2). \quad (31)$$

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

$$\varphi(w') - \varphi(w) = (w' - w) \partial_w \varphi(w) + \frac{1}{2} (w' - w)^2 \partial_w^2 \varphi(\tilde{w}) \quad (32)$$

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

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

Therefore the approximation of the interaction integral in (25) reads

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \left\langle \int_{I^2} (w' - w) \partial_w \varphi(w) f(w) f(w_*) dw dw_* \right. \\ & \left. + \frac{1}{2} (w' - w)^2 \partial_w^2 \varphi(w) f(w) f(w_*) dw dw_* \right\rangle + R(\varepsilon). \end{aligned} \quad (33)$$

The term $R(\varepsilon)$ indicates the remainder of the Taylor expansion and is such that

$$R(\varepsilon) = \frac{1}{2\varepsilon} \left\langle \int_{I^2} (w' - w)^2 (\partial_w^2 \varphi(\tilde{w}) - \partial_w^2 \varphi(w)) f(w) f(w_*) dw dw_* \right\rangle. \quad (34)$$

Under suitable assumptions on the function space of φ and ξ the remainder converges to zero as soon as $\varepsilon \rightarrow 0$ (see [98]). Thanks to (31) the limit operator of (33) is the following

$$\int_{I^2} \left(P(w, w_*)(w_* - w) + \frac{2}{\kappa} K(w, w_*) \right) \partial_w \varphi(w) f(w) f(w_*) dw dw_* \\ + \frac{\sigma^2}{2} \int_I D(w)^2 \partial_w^2 \varphi(w) f(w) dw.$$

Integrating back by parts the last expression, and supposing that the border terms vanish, we obtain the following Fokker–Planck equation

$$\frac{\partial}{\partial t} f + \frac{\partial}{\partial w} \mathcal{P}[f](w) f(w) + \frac{\partial}{\partial w} \mathcal{K}[f](w) f(w) dv = \frac{\sigma^2}{2} \frac{\partial^2}{\partial w^2} (D(w)^2 f(w)), \quad (35)$$

where

$$\mathcal{P}[f](w) = \int_I P(w, v)(v - w) f(v) dv, \\ \mathcal{K}[f](w) = \frac{2}{\kappa} \int_I K(w, v) f(v) dv = \frac{1}{\kappa} ((w_d - w) + (w_d - m)).$$

As usual, $m(t) = \int_I w f(w, t) dw$ indicates the mean opinion.

2.1.5 Stationary Solutions

One of the advantages of the Fokker–planck description is related to the possibility to identify analytical steady states. In this section we will look for steady solutions of the Fokker–Planck model (35), for particular choices of the microscopic interaction of the Boltzmann dynamics.

The stationary solutions, say $f_\infty(w)$, of (35) satisfy the equation

$$\frac{\partial}{\partial w} \mathcal{P}[f](w) f(w) + \frac{\partial}{\partial w} \mathcal{K}[f](w) f(w) dv = \frac{\sigma^2}{2} \frac{\partial^2}{\partial w^2} (D(w)^2 f(w)). \quad (36)$$

As shown in [7, 9, 98], equation (36) can be analytically solved under some simplifications. In general solutions to (36) satisfy the ordinary differential equation

$$\frac{df}{dw} = \left(\frac{2}{\sigma^2} \frac{\mathcal{P}[f](w) + \mathcal{K}[f](w)}{D(w)^2} - 2 \frac{D'(w)}{D(w)} \right) f. \quad (37)$$

Thus

$$f(w) = \frac{C_0}{D(w)^2} \exp \left\{ \frac{2}{\sigma^2} \int^w \left(\frac{\mathcal{P}[f](v) + \mathcal{K}[f](v)}{D(v)^2} \right) dv \right\}, \quad (38)$$

where C_0 is a normalizing constant.

Let us consider the simpler case in which $P(w, v) = 1$. Then the average opinion $m(t)$ evolves according to

$$m(t) = (1 - e^{-2/\kappa t}) w_d + e^{-2/\kappa t} m(0), \quad (39)$$

which is obtained from the scaled equation (30) through the quasi-invariant opinion limit (cf. also equation (28) for a comparison).

In absence of control, i.e. for $\kappa \rightarrow \infty$, the mean opinion is conserved, and the steady solutions of (35) satisfy the differential equation [98]

$$\partial_w(D(w)^2 f) = \frac{2}{\sigma^2} (m - w) f. \quad (40)$$

In presence of the control the mean opinion is in general not conserved in time, even if, from (39), it is clear that $m(t)$ converges exponentially in time to w_d . Consequently

$$\partial_w(D(w)^2 f) = \frac{2}{\sigma^2} \left(1 + \frac{1}{\kappa}\right) (w_d - w) f. \quad (41)$$

Let us consider as diffusion function $D(w) = (1 - w^2)$. Therefore the solution of (40) takes the form

$$\begin{aligned} f_\infty(w) &= \frac{C_{m,\sigma}}{(1 - w^2)^2} \left(\frac{1+w}{1-w}\right)^{m/(2\sigma^2)} \exp\left\{-\frac{1-mw}{\sigma^2(1-w^2)}\right\} \\ &= \frac{C_{m,\sigma}}{(1 - w^2)^2} S_{m,\sigma^2}(w), \end{aligned} \quad (42)$$

where $C_{m,\sigma}$ is such that the mass of f_∞ is equal to one. This solution is regular, and thanks to the presence of the exponential term $f(\pm 1) = 0$. Moreover, due to the general non symmetry of f , the initial opinion distribution reflects on the steady state through the mean opinion. The dependence on κ can be rendered explicit. It gives

$$f_\infty^\kappa(w) = \frac{C_{w_d,\sigma,\kappa}}{(1 - w^2)^2} (S_{w_d,\sigma}(w))^{1+1/\kappa}, \quad (43)$$

with $C_{w_d,\sigma,\kappa}$ the normalization constant.

We plot in Figure 1 the steady profile f_∞ and f_∞^κ for different choice 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 $\kappa \rightarrow \infty$ the steady profile of (41) approaches the one of (40). On the other hand small values of κ give the desired distribution concentrated around w_d . It is remarkable that in general we can not switch from f_∞ to f_∞^κ only acting on the parameter κ , since the memory on the initial average opinion is lost for any $\kappa > 0$. We refer to [6, 98] for further discussion about stationary solutions of (36).

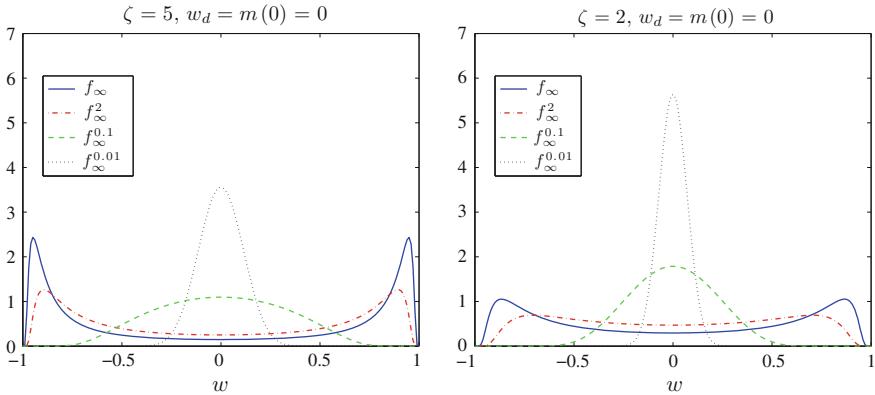


Fig. 1 Continuous line and dashed lines represent respectively the steady solutions f_∞ and f_∞^κ . On the left $w_d = m(0) = 0$ with diffusion parameter $\zeta = \sigma^2 = 5$, on the right $w_d = m(0) = 0$ with diffusion parameter $\zeta = \sigma^2 = 2$. In both cases the steady solution changes from a bimodal distribution to an unimodal distribution around w_d .

2.1.6 Numerical Tests

Our goal is to investigate the action of the control dynamics at the mesoscopic level. We solve directly the kinetic equation (25) obeying the binary interaction (31), for small value of the scale parameter $\varepsilon > 0$. We perform the numerical simulations using the Monte Carlo methods developed in [5, 89].

Sznajd's Model

Our first example refers to the mean-field Sznajd's model [11, 97]

$$\partial_t f = \gamma \partial_w (w(1-w^2)f), \quad (44)$$

corresponding to equation (35) in the uncontrolled case without diffusion. It is obtained choosing $P(w, v) = 1 - w^2$ and assuming that the mean opinion $m(t)$ is always zero. In [11] authors showed for $\gamma = 1$ *concentration* of the profile around zero, and conversely for $\gamma = -1$ a *separation* phenomena, namely concentration around $w = 1$ and $w = -1$, by showing that explicit solutions are computable. We approximate the mean-field dynamics in the *separation* case, $\gamma = -1$, through the binary interaction (31), sampling $N_s = 1 \times 10^5$ agents, with scaling parameter $\varepsilon = 0.005$. In Figure 2 we simulate the evolution of $f(w, t)$ in the time interval $[0, 2]$, starting from the uniform distribution on I , $f_0(w) = 1/2$, in three different cases: uncontrolled ($\kappa = \infty$), mild control ($\kappa = 1$) and strong control ($\kappa = 0.1$). In the controlled cases the distribution is forced to converge to the desired state $w_d = 0$.

Bounded Confidence Model

We consider now the *bounded confidence model* introduced in [72], where every agent interacts only within a certain level of confidence. This can be modelled through the potential function

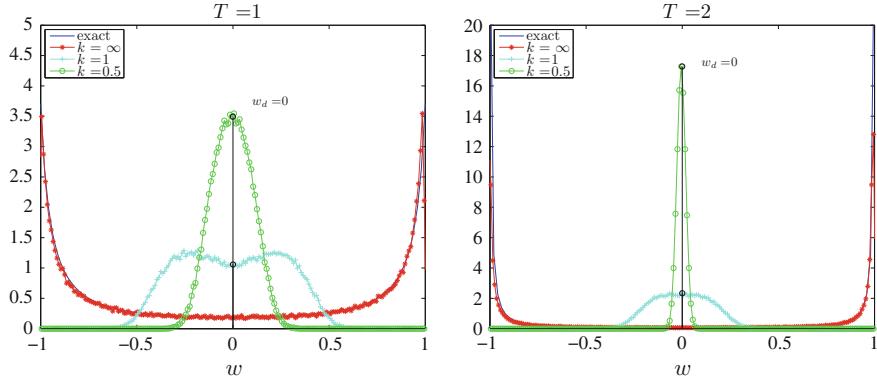


Fig. 2 Solution profiles at time $T = 1$, and $T = 2$, for uncontrolled $\kappa = \infty$, mildly controlled case $\kappa = 1$, strong controlled case $\kappa = 0.1$. Desired state is set to $w_d = 0$.

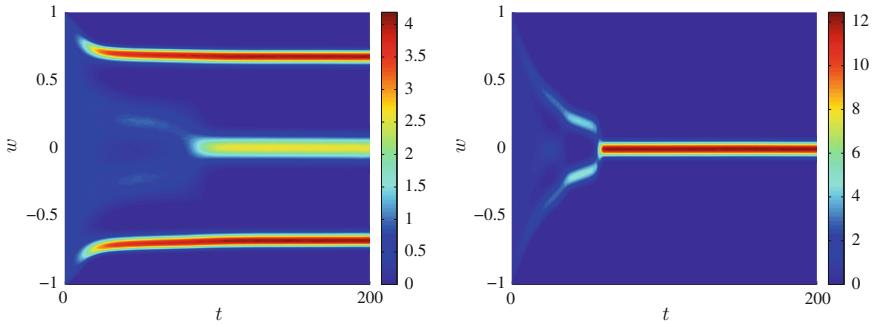


Fig. 3 On the right the penalization of the control parameter is $\kappa = 5 \times 10^3$ on the left $\kappa = 0.001$. Evolution of the kinetic density, using $N_s = 2 \times 10^5$ sample particles on a 200×400 grid. Binary interactions (31) performed with $\varepsilon = 0.01$ and $\varsigma = 0.01$, $\Delta = 0.2$.

$$P(w, v) = \chi(|w - v| \leq \Delta), \quad \Delta < 2.$$

In Figure 3, we simulate the dynamics of the agents starting from a uniform distribution of the opinions on the interval $I = [-1, 1]$. The binary interaction (31) refers to a diffusion parameter $\sigma = 0.01$ and $\varepsilon = 0.05$. Here $N_s = 2 \times 10^5$. The bounded confidence parameter is $\Delta = 0.2$, and we consider both cases (without control and with control), letting the system evolve in the time interval $[0, T]$, with $T = 200$. The figure to the left refers to the uncontrolled case, where three mainstream opinions emerge. On the right the presence of the control with $\kappa = 5$ leads the opinions to concentrate around the desired opinion $w_d = 0$.

2.2 Control Through Leadership

Several studies have been recently focused on the control of a large population through the action of a small portion of individuals, typically identified as leaders [3, 4, 24, 62]. In this section we are interested in the opinion formation process of a followers' population steered by the action of a leaders' group. At a microscopic level we suppose to have a population of N_F followers and N_L leaders. Their dynamics is modelled as follows

$$\dot{w}_i = \frac{1}{N_F} \sum_{j=1}^{N_F} P(w_i, w_j) (w_j - w_i) + \frac{1}{N_L} \sum_{h=1}^{N_L} S(w_i, v_h) (v_h - w_i), \quad w_i(0) = w_{i,0}, \quad (45)$$

$$\dot{v}_k = \frac{1}{N_L} \sum_{h=1}^{N_L} R(v_k, v_h) (v_h - v_k) + u, \quad v_k(0) = v_{k,0}, \quad (46)$$

where $w_i, v_k \in I$, $I = [-1, 1]$ for all $i = 1, \dots, N_F$ and $k = 1, \dots, N_L$ are the followers' and leaders' opinions. As in the previous section, $P(\cdot, \cdot)$, $S(\cdot, \cdot)$ and $R(\cdot, \cdot)$ are given *compromise functions*, measuring the relative importance of the interacting agent in the consensus dynamics. Leaders' strategy is driven by a suitable control u , which minimizes the functional

$$J(u) = \frac{1}{2} \int_0^T \left(\frac{\psi}{N_L} \sum_{h=1}^{N_L} (v_h - w_d)^2 + \frac{\mu}{N_L} \sum_{h=1}^{N_L} (v_h - m_F)^2 \right) dt + \frac{\kappa}{2} \int_0^T u^2 dt, \quad (47)$$

where T represents the final time horizon, w_d is the desired opinion and m_F is the average opinion of the followers group at time $t \geq 0$, and $\psi, \mu > 0$ are such that $\psi + \mu = 1$. Therefore, leaders' behavior is driven by a suitable control strategy based on the interplay between the desire to force followers towards a given state, *radical behavior* ($\psi \approx 1$), and the necessity to keep a position close to the mean opinion of the followers in order to influence them *populistic behavior* ($\mu \approx 1$).

Note that, since the optimal control problem acts only over the leader dynamics we can approximate its solution by a model predictive control approximation as in Section 2.1. Next we can build the corresponding constrained binary Boltzmann dynamics following [7].

2.2.1 Boltzmann Constrained Dynamics

To derive the system of kinetic equation we introduce a density distribution of followers $f_F(w, t)$ and leaders $f_L(v, t)$ depending on the opinion variables $w, v \in I$ and time $t \geq 0$, see [7, 59]. It is assumed that the densities of the followers and the leaders satisfy is

$$\int_I f_F(w, t) dw = 1, \quad \int_I f_L(v, t) dv = \rho \leq 1.$$

The kinetic model can be derived by considering the change in time of $f_F(w, t)$ and $f_L(v, t)$ depending on the interactions with the other individuals and on the leaders' strategy. This change depends on the balance between the gain and loss due to the binary interactions. Starting by the pair of opinions (w, w_*) and (v, v_*) , respectively the opinions of two followers and two leaders, the post-interaction opinions are computed according to three dynamics: *a*) the interaction between two followers; *b*) the interaction between a follower and a leader; *c*) the interaction between two leaders.

- a)* We assume that the opinions (w', w'_*) in the follower-follower interactions obey to the rule

$$\begin{cases} w' = w + \eta P(w, w_*)(w_* - w) + \xi D_F(w), \\ w'_* = w_* + \eta P(w_*, w)(w - w_*) + \xi_* D_F(w_*), \end{cases} \quad (48)$$

where as usual $P(\cdot, \cdot)$ is the compromise function, and the diffusion variables ξ, ξ_* are realizations of a random variable with zero mean, finite variance ς_F^2 . The noise influence is weighted by the function $D_F(\cdot)$, representing the local relevance of diffusion for a given opinion, and such that $0 \leq D_F(\cdot) \leq 1$.

- b)* The leader-follower interaction is described for every agent from the leaders group. Since the leader do not change opinion, we have

$$\begin{cases} w'' = w + \eta S(w, v)(v - w) + \zeta D_{FL}(w) \\ v'' = v \end{cases} \quad (49)$$

where $S(\cdot, \cdot)$ is the communication function and ζ a random variable with zero mean and finite variance ς_{FL}^2 , weighted again by the function $D_{FL}(\cdot)$.

- c)* Finally, the post-interaction opinions (v', v'_*) of two leaders are given by

$$\begin{cases} v' = v + \eta R(v, v_*)(v_* - v) + \eta U(v, v_*; m_F) + \theta D_L(v) \\ v'_* = v_* + \eta R(v_*, v)(v - v_*) + \eta U(v, v_*; m_F) + \theta_* D_L(v_*), \end{cases} \quad (50)$$

where $R(\cdot, \cdot)$ is the compromise function and, similar to the previous dynamics, θ, θ_* are random variables with zero mean and finite variance ς_L^2 , weighted by $D_L(\cdot)$. Moreover the leaders' dynamics include the feedback control, derived from (47) with the same approach of Section 2.1.1. In this case the feedback control accounts for the average values of the followers' opinion

$$m_F(t) = \int_I w f_F(w, t) dw, \quad (51)$$

and it is defined as

$$\eta U(v, v_*; m_F) = \beta [K(v, v_*; m_F) + \eta H(v, v_*)]. \quad (52)$$

In (52), β has the same form of (21) and

$$K(v, v_*; m_F) = \frac{\psi}{2} ((w_d - v) + (w_d - v_*)) + \frac{\mu}{2} ((m_F - v) + (m_F - v_*)), \quad (53)$$

$$H(v, v_*) = \frac{1}{2}(R(v, v_*) - R(v_*, v))(v - v_*). \quad (54)$$

Note that the control term, $K(v, v_*; m_F)$ depends on two contributions, a steering force towards the desired state w_d and one towards the average opinion of the followers m_F , weighted respectively by the parameters ψ and μ , such that $\psi + \mu = 1$.

2.2.2 Boltzmann–Type Modeling

Following [89], for a suitable choice of test functions φ we can describe the evolution of $f_F(w, t)$ and $f_L(t)$ via a system of integro-differential equations of Boltzmann type

$$\begin{cases} \frac{d}{dt} \int_I \varphi(w) f_F(w, t) dw = (Q_F(f_F, f_F), \varphi) + (Q_{FL}(f_F, f_L), \varphi), \\ \frac{d}{dt} \int_I \varphi(v) f_L(v, t) dv = (Q_L(f_L, f_L), \varphi). \end{cases} \quad (55)$$

The operators Q_F , Q_{FL} and Q_L account for the binary exchange of opinions. Under the assumption that the interaction parameters are such that $|w'|, |w''|, |v'| \leq 1$ the action of the Boltzmann operators on a (smooth) function φ can be written as

$$(Q_F(f_F, f_F), \varphi) = \lambda_F \left\langle \int_{I^2} (\varphi(w') - \varphi(w)) f_F(w, t) f_F(v, t) dw dv \right\rangle, \quad (56)$$

$$(Q_{FL}(f_F, f_L), \varphi) = \lambda_{FL} \left\langle \int_{I^2} (\varphi(w'') - \varphi(w)) f_F(w, t) f_L(v_*, t) dw dv_* \right\rangle, \quad (57)$$

$$(Q_L(f_L, f_L), \varphi) = \lambda_L \left\langle \int_{I^2} (\varphi(v') - \varphi(v)) f_L(v, t) f_L(v_*, t) dv dv_* \right\rangle, \quad (58)$$

where $\lambda_F, \lambda_{FL}, \lambda_L > 0$ are constant relaxation rates and, as before, $\langle \cdot \rangle$ denotes the expectation taken with respect to the random variables characterizing the noise terms.

To study the evolution of the average opinions, we can take $\varphi(w) = w$ in (55). In general this leads to a complicated nonlinear system [7], however in the simplified situation of P and R symmetric and $S \equiv 1$ we obtain the following closed system of differential equations for the mean opinions

$$\begin{cases} \frac{d}{dt}m_L(t) = \tilde{\eta}_L\psi\beta(w_d - m_L(t)) + \tilde{\eta}_L\mu\beta(m_F(t) - m_L(t)) \\ \frac{d}{dt}m_F(t) = \tilde{\eta}_{FL}\alpha(m_L(t) - m_F(t)), \end{cases} \quad (59)$$

where we introduced the notations $\tilde{\eta}_L = \rho\eta_L$, $\tilde{\eta}_{FL} = \rho\eta_{FL}$ and $m_L(t) = \frac{1}{\rho} \int_I v f_L(v, t) dv$.

Straightforward computations show that the exact solution of the above system has the following structure

$$\begin{cases} m_L(t) = C_1 e^{-|\lambda_1|t} + C_2 e^{-|\lambda_2|t} + w_d \\ m_F(t) = C_1 \left(1 + \frac{\lambda_1}{\beta\mu\tilde{\eta}_L}\right) e^{-|\lambda_1|t} + C_2 \left(1 + \frac{\lambda_2}{\beta\mu\tilde{\eta}_L}\right) e^{-|\lambda_2|t} + w_d \end{cases} \quad (60)$$

where C_1, C_2 depend on the initial data $m_F(0), m_L(0)$ in the following way

$$\begin{aligned} C_1 &= -\frac{1}{\lambda_1 - \lambda_2} ((\beta\tilde{\eta}_L m_L(0) + \lambda_2)m_L(0) - \mu\beta\tilde{\eta}_L m_F(0) - (\lambda_2 + \beta\tilde{\eta}_L\psi)w_d) \\ C_2 &= \frac{1}{\lambda_1 - \lambda_2} ((\beta\tilde{\eta}_L m_L(0) + \lambda_1)m_L(0) - \mu\beta\tilde{\eta}_L m_F(0) - (\lambda_1 + \beta\tilde{\eta}_L\psi)w_d) \end{aligned}$$

with

$$\lambda_{1,2} = -\frac{1}{2}(\alpha\tilde{\eta}_{FL} + \beta\tilde{\eta}_L) \pm \frac{1}{2}\sqrt{(\alpha\tilde{\eta}_{FL} + \beta\tilde{\eta}_L)^2 - 4\psi\alpha\beta\tilde{\eta}_L\tilde{\eta}_{FL}}.$$

Note that $\lambda_{1,2}$ are always negative, this assures that the contribution of the initial averages, $m_L(0), m_F(0)$, vanishes as soon as time increases and the mean opinions of leaders and followers converge towards the desired state w_d . Moreover, in absence of diffusion, it can be shown that the corresponding variance vanishes [7], i.e. under the above assumptions the steady state solutions have the form of a Dirac delta centered in the target opinion w_d .

2.2.3 Fokker-Planck Modeling

Once more, the study of the large-time behavior of the kinetic equation (55) will take advantage by passing to a Fokker-Planck description. Therefore, similarly to Section 2.1, we consider the quasi-invariant opinion limit [7, 89, 98], introducing the parameter $\varepsilon > 0$, and scaling the quantities in the binary interaction

$$\begin{aligned} \eta &= \varepsilon, & \varsigma_F &= \sqrt{\varepsilon}\sigma_F, & \varsigma_L &= \sqrt{\varepsilon}\sigma_L, & \varsigma_{FL} &= \sqrt{\varepsilon}\sigma_{FL}, \\ \lambda_F &= \frac{1}{c_F\varepsilon}, & \lambda_{FL} &= \frac{1}{c_{FL}\varepsilon}, & \lambda_L &= \frac{1}{c_L\varepsilon}, & \beta &= \frac{2\varepsilon}{\kappa + 2\varepsilon}. \end{aligned} \quad (61)$$

The scaled equation (55) in the *quasi-invariant opinion limit* is well approximated by a Fokker-Planck equation for the followers' opinion distribution

$$\frac{\partial f_F}{\partial t} + \frac{\partial}{\partial w} ((\mathcal{P}[f_F](w) + \mathcal{S}[f_L](w)) f_F(w)) = \frac{\partial^2}{\partial w^2} (\mathcal{D}_F[f_F, f_L](w) f_F(w)), \quad (62)$$

where

$$\begin{aligned} \mathcal{P}[f_F](w) &= \frac{1}{c_F} \int_I P(w, w_*)(w_* - w) f_F(w_*) dw_*, \\ \mathcal{S}[f_L](w) &= \frac{1}{c_{FL}} \int_I S(w, v_*)(v_* - w) f_L(v_*) dv_*, \\ \mathcal{D}_F[f_F, f_L](w) &= \frac{\sigma_F^2}{2c_F} D_F(w)^2 + \frac{\sigma_{FL}^2 \rho}{2c_{FL}} D_{FL}(w)^2, \end{aligned}$$

and an equivalent Fokker-Planck equation for the leaders' opinion distribution

$$\frac{\partial f_L}{\partial t} + \frac{\partial}{\partial v} ((\mathcal{R}[f_L](v) + \mathcal{K}[f_L, f_F](v)) f_L(v)) = \frac{\partial^2}{\partial v^2} (\mathcal{D}_L[f_L](v) f_L(v)), \quad (63)$$

where

$$\begin{aligned} \mathcal{R}[f_L](v) &= \frac{\rho}{c_L} \int_I R(v, v_*)(v_* - v) f_L(v_*) dv_*, & \mathcal{D}_L[f_L](v) &= \frac{\sigma_L^2 \rho}{2c_L} D_L^2(v), \\ \mathcal{K}[f_L, f_F](\tilde{w}) &= \frac{\psi}{\kappa c_L} (v + m_L(t) - 2w_d) + \frac{\mu}{\kappa c_L} (v + m_L(t) - 2m_F(t)). \end{aligned}$$

In some cases it is possible to recover explicitly the stationary states of the Fokker-Planck system (62) and (63). In the simplified case where every interaction function is constant and unitary, i.e. $P \equiv S \equiv R \equiv 1$, and $D_F(w) = D_L(w) = D_{FL}(w) = 1 - w^2$, we have

$$f_{F,\infty} = \frac{a_F}{(1-w^2)^2} \exp \left\{ -\frac{2}{b_F} \int_0^w \frac{z - w_d}{(1-z^2)^2} dz \right\}, \quad b_F = \frac{\sigma_F^2 c_{FL} + \sigma_{FL}^2 c_F \rho}{c_{FL} + c_F \rho} \quad (64)$$

$$f_{L,\infty} = \frac{a_L}{(1-\tilde{w}^2)^2} \exp \left\{ -\frac{2}{b_L} \int_0^{\tilde{w}} \left(\frac{z - w_d}{(1-z^2)^2} \right) dz \right\}, \quad b_L = \frac{\sigma_L^2 \rho \kappa}{2c_L(\psi + \mu)}, \quad (65)$$

where a_F, a_L are suitable normalization constants. We refer to [7] for further details.

Table 1 Computational parameters for the different test cases.

Test	$S(\cdot, \cdot)$	c_F	\hat{c}_{FL}	\hat{c}_L	ρ	ψ	w_d					
#1	eq. (66)	1	0.1	0.1	0.05	0.5	0.5					
	$S(\cdot, \cdot)$	c_F	\hat{c}_{FL_1}	\hat{c}_{L_1}	ρ_1	ψ_1	w_{d_1}	\hat{c}_{FL_2}	\hat{c}_{L_2}	ρ_2	ψ_2	w_{d_2}
#2	1	1	0.1	0.1	0.05	eq. (69)	0.5	1	0.1	0.05	eq. (69)	-0.5

2.2.4 Numerical Experiments

In this section we present some numerical results concerning the simulation of the Boltzmann type control model (55). All the results have been computed using the Monte Carlo method for the Boltzmann model developed in [5] in the Fokker-Planck regime $\varepsilon = 0.01$ under the scaling (61).

In the numerical tests we assume that the $\rho_L = 0.05$, (five per cent of the population is composed by opinion leaders [59]). Note that, for clarity, in all figures the leaders' profiles have been scaled by a factor 10. The random diffusion effects have been computed in the case of a uniform random variable with $\sigma_F^2 = \sigma_{FL}^2 = \sigma_L^2 = 0.01$. It is easy to check that the above choice preserves the bounds in the numerical simulations. First we present some test cases with a single population of leaders. Then we consider the case of multiple populations of leaders with different time-dependent strategies. This leads to more realistic applications of our arguments, introducing the concept of competition between the leaders. For the sake of simplicity, the interaction functions $P(\cdot, \cdot)$ and $R(\cdot, \cdot)$ are assumed to be constant. The remaining computational parameters have been summarized in Table 1.

Test 1. Leaders Driving Followers

In the first test case we consider the system of Boltzmann equations (55) with a single population of leaders driving followers.

We assume the initial distributions $f_F \sim U([-1, -0.5])$ and $f_L \sim N(w_d, 0.05)$, where $U(\cdot)$ and $N(\cdot, \cdot)$ denote the uniform and the normal distributions respectively. We consider constant interaction functions $P(\cdot, \cdot)$ and $R(\cdot, \cdot)$ and a bounded confidence-type function for the leader-follower interactions

$$S(w, v) = \chi(|w - v| \leq \Delta), \quad (66)$$

with $\Delta = 0.5$. Other parameters are defined in Table 1, where we used the compact notations $\hat{c}_{FL} = c_{FL}/\rho$ and $\hat{c}_L = c_L/\rho$.

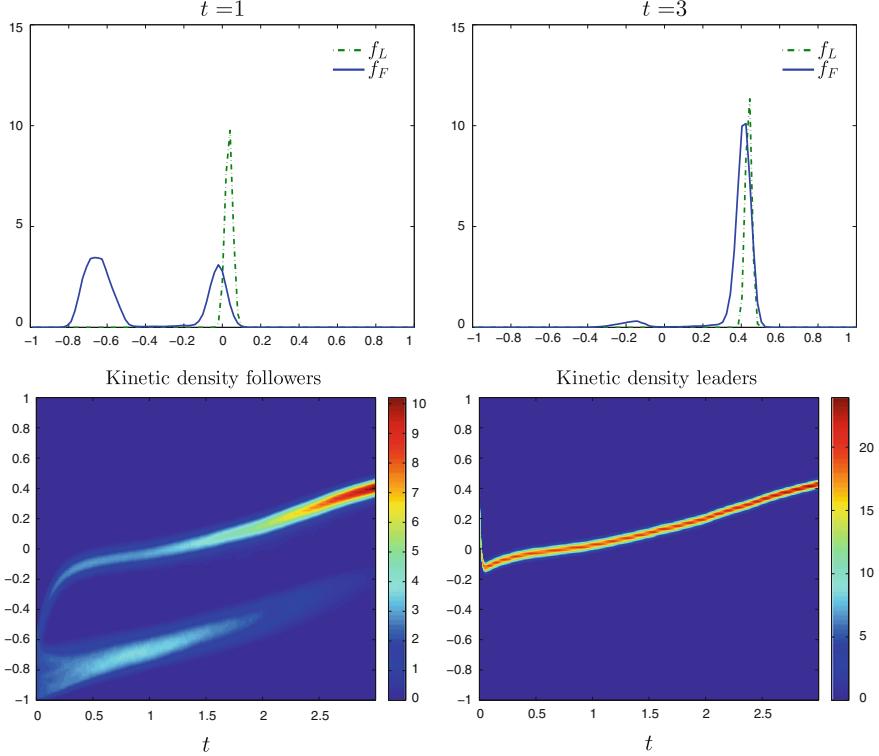


Fig. 4 Test #1: Kinetic densities at different times for a single population of leaders with bounded confidence interaction (top row). Kinetic densities evolution over the time interval $[0, 3]$ (bottom row).

In Figure 4 we report the evolution, over the time interval $[0, 3]$, of the kinetic densities $f_F(w, t)$ and $f_L(v, t)$. The numerical experiment shows that the optimal control problem is able to generate a non monotone behavior of $m_L(t)$, resulting from the combined leaders' strategy of a populists and radical behavior. In an electoral context, this is a characteristic which can be found in populist radical parties, which typically include non-populist ideas and their leadership generates through a dense network of radical movements [85].

Test 2. The Case of Competing Multi-leaders Populations

When more than one population of leaders is present, each one with a different strategy, we describe the evolution of the kinetic density of the system through a Boltzmann approach. Let $M > 0$ be the number of families of leaders, each of them described by the density f_{L_j} , $j = 1, \dots, M$ such that

$$\int_I f_{L_j}(v) dv = \rho_j. \quad (67)$$

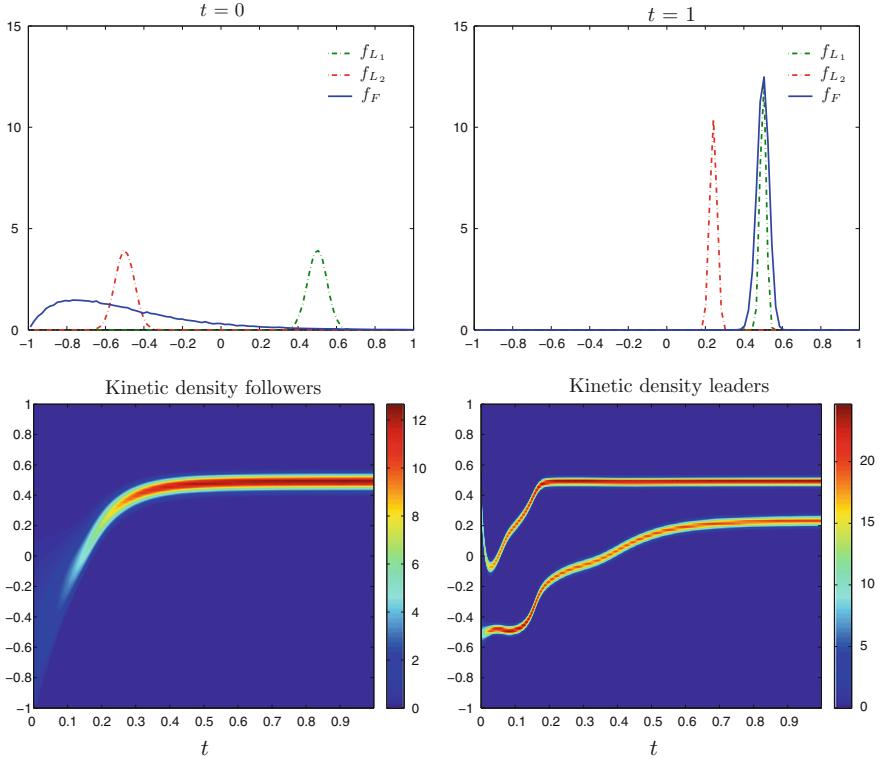


Fig. 5 Test #2: Kinetic densities at different times for for a two populations of leaders model with time dependent strategies (top row). Kinetic density evolution over the time interval $[0, 1]$ (bottom row).

If a unique population of followers is present, with density f_F , a follower interacts both with the others agents from the same population and with every leader of each j -th family. Given a suitable test function φ the evolution of the densities is given by the system of Boltzmann equations

$$\begin{cases} \frac{d}{dt} \int_I \varphi(w) f_F(w, t) dw = (Q_F(f_F, f_F), \varphi) + \sum_{k=1}^M (Q_{FL}(f_{L_k}, f_F), \varphi), \\ \frac{d}{dt} \int_I \varphi(\tilde{w}) f_{L_j}(\tilde{w}, t) d\tilde{w} = (Q_L(f_{L_j}, f_{L_j}), \varphi), \quad j = 1, \dots, M. \end{cases} \quad (68)$$

By assuming that the leaders aim at minimizing cost functionals of the type (47), the differences consist in two factors: in the target opinions w_{d_j} and in the leaders' attitude towards a radical ($\psi_j \approx 1$) or populist strategy ($\mu_j \approx 1$).

To include *competition* between different populations, we introduce time-dependent coefficients in the leaders' strategies. This approach leads to the con-

cept of *adaptive strategy* for every family of leaders $j = 1, \dots, M$. Thus we assume that coefficients ψ and μ which appear into the functional now evolve in time and are defined for each $t \in [0, T]$ as

$$\psi_j(t) = \frac{1}{2} \int_{w_{d_j}-\delta}^{w_{d_j}+\delta} f_F(w) dw + \frac{1}{2} \int_{m_{L_j}-\bar{\delta}}^{m_{L_j}+\bar{\delta}} f_F(w) dw, \quad \mu_j(t) = 1 - \psi_j(t) \quad (69)$$

where both $\delta, \bar{\delta} \in [0, 1]$ are fixed and m_{L_j} is the average opinion of the j th population of leaders. The introduced choice of coefficients is equivalent to consider a competition between the populations of leaders, where each leader try to adapt its populist or radical attitude accordingly to the success of the strategy. Note also that the success of the strategy is based on the local perception of the followers.

In the numerical experiments reported in Figure 5 we take into account two populations of leaders, initially normally distributed with mean values w_{d_1} and w_{d_2} and parameters $\delta = \bar{\delta} = 0.5$, respectively, and a single population of followers, represented by a skewed distribution $f_F \sim \Gamma(2, \frac{1}{4})$ over the interval $[-1, 1]$, where $\Gamma(\cdot, \cdot)$ is the Gamma distribution. Here the frequencies of interactions are assumed to be unbalanced since $\hat{c}_{FL_1} = 0.1$ and $\hat{c}_{FL_2} = 1$. In the test case we assume that the followers group has an initial natural inclination for the position represented by one leader but, thanks to communication strategies pursued by the minority leader, it is driven to different positions (see Figures 5). In a bipolar electoral context, an example of the described behavior would consist of a better use of the media in a coalition with respect to the opponents.

3 Multivariate Models

In several recent works additional variables have been introduced quantifying relevant indicators for the spreading of opinions [9, 10, 31, 53, 58, 91]. In this class of models the opinion dynamics depends on an additional parameter, continuous or discrete, which influences the binary exchanges. We present in this section two kinetic multivariate models. The first takes into account a continuous variable called *conviction* representing the strength of individuals in pursuing their opinions. Afterwards we develop a model for the dynamics of opinions in large evolving networks where the *number of connections* of each individuals, a discrete variable, influences the dynamics.

3.1 The Role of Conviction

Resembling the model for wealth exchange in a multi-agent society introduced in [38], this new model has an additional parameter to quantify the personal *conviction*,

representing a measure of the influencing ability of individuals [31]. Individuals with high conviction are resistant to change opinion, and have a prominent role in attracting other individuals towards their opinions. In this sense, individuals with high conviction play the role of leaders [59].

The goal is to study the evolution of a multi-agent system characterized by two variables, representing conviction and opinion, where the way in which conviction is formed is independent of the personal opinion. Then, the (personal) conviction parameter will enter into the microscopic binary interactions for opinion formation considered in Section 1.2, to modify them in the compromise and self-thinking terms. A typical and natural assumption is that high conviction could act on the interaction process both to reduce the personal propensity to compromise, and to reduce the self-thinking. Numerical investigation shows that the role of the additional conviction variable is to bring the system towards a steady distribution in which there is formation of clusters even in absence of bounded confidence hypotheses [18–20, 72].

3.1.1 The Formation of Conviction

Let us briefly summarize the key points at the basis of the model for conviction [31]. Each variation is interpreted as an interaction where a fraction of the conviction of the individual is lost by virtue of afterthoughts and insecurities, while at the same time the individual can absorb a certain amount of conviction through the information achieved from the external background (the surrounding environment). In this approach, the conviction of the individual is quantified in terms of a scalar parameter x , ranging from zero to infinity. Denoting with $z \geq 0$ the degree of conviction achieved from the background, it is assumed that the new amount of conviction in a single interaction can be computed as

$$x^* = (1 - \lambda(x))x + \lambda_B(x)z + \vartheta H(x). \quad (70)$$

In (70) the functions $\lambda = \lambda(x)$ and $\lambda_B = \lambda_B(x)$ quantify, respectively, the personal amounts of insecurity and willingness to be convinced by others, while ϑ is a random parameter which takes into account the possible unpredictable modifications of the conviction process. We will in general fix the mean value of ϑ equal to zero. Last, $H(\cdot)$ will denote an increasing function of conviction. The typical choice is to take $H(x) = x^\nu$, with $0 < \nu \leq 1$. Since some insecurity is always present, and at the same time it can not exceed a certain amount of the total conviction, it is assumed that $\lambda_- \leq \lambda(x) \leq \lambda_+$, where $\lambda_- > 0$, and $\lambda_+ < 1$. Likewise, we will assume an upper bound for the willingness to be convinced by the environment. Then, $0 \leq \lambda_B(x) \leq \bar{\lambda}$, where $\bar{\lambda} < 1$. Lastly, the random part is chosen to satisfy the lower bound $\vartheta \geq -(1 - \lambda_+)$. By these assumptions, it is assured that the post-interaction value x^* of the conviction is nonnegative.

Let $C(z)$, $z \geq 0$ denote the probability distribution of degree of conviction of the (fixed) background. We will suppose that $C(z)$ has a bounded mean, so that

$$\int_{\mathbb{R}_+} C(z) dz = 1; \quad \int_{\mathbb{R}_+} z C(z) dz = M \quad (71)$$

We note that the distribution of the background will induce a certain policy of acquisition of conviction. This aspect has been discussed in [90], from which we extract the example that follow. Let us assume that the background is a random variable uniformly distributed on the interval $(0, a)$, where $a > 0$ is a fixed constant. If we choose for simplicity $\lambda(x) = \lambda_B(x) = \bar{\lambda}$, and the individual has a degree of conviction $x > a$, in absence of randomness the interaction will always produce a value $x^* \leq x$, namely a partial decrease of conviction. In this case, in fact, the process of insecurity in an individual with high conviction can not be restored by interaction with the environment.

The study of the time-evolution of the distribution of conviction produced by binary interactions of type (70) can be obtained by resorting to kinetic collision-like models [89]. Let $F = F(x, t)$ the density of agents which at time $t > 0$ are represented by their conviction $x \in \mathbb{R}_+$. Then, the time evolution of $F(x, t)$ obeys to a Boltzmann-like equation. This equation is usually written in weak form. It corresponds to say that the solution $F(x, t)$ satisfies, for all smooth functions $\varphi(x)$ (the observable quantities)

$$\frac{d}{dt} \int_{\mathbb{R}_+} F(x, t) \varphi(x) dx = \left\langle \int_{\mathbb{R}_+^2} (\varphi(x^*) - \varphi(x)) F(x, t) C(z) dx dz \right\rangle, \quad (72)$$

where x^* is the post-interaction conviction and $\langle \cdot \rangle$ denotes the expectation with respect to the random parameter ϑ introduced in (70). Through the techniques analyzed in the previous sections of this work we can derive the asymptotic solution of the Fokker-Planck equation which follows from (72) in the limit $\varepsilon \rightarrow 0$.

If $\lambda(x) = \lambda$ and $\lambda_B(x) = \lambda_B$ we get the explicit form of the steady distribution of conviction [31, 89]. We will present two realizations of the asymptotic profile, that enlighten the consequences of the choice of a particular function $H(\cdot)$. First, let us consider the case in which $H(x) = x$. In this case, the Fokker-Planck equation coincides with the one obtained in [44], related to the steady distribution of wealth in a multi-agent market economy. One obtains

$$G_\infty(x) = \frac{G_0}{x^{2+2\lambda/\mu}} \exp \left\{ -\frac{2\lambda_B M}{\mu x} \right\}, \quad (73)$$

where the constant G_0 is chosen to fix the total mass of $G_\infty(x)$ equal to one. Note that the steady profile is heavy tailed, and the size of the polynomial tails is related to both λ and σ . Hence, the percentage of individuals with high conviction is decreasing as soon as the parameter λ of insecurity is increasing, and/or the parameter of self-thinking is decreasing. It is moreover interesting to note that the size of the parameter λ_B is important only in the first part of the x -axis, and contributes to determine the size of the number of undecided.

The second case refers to the choice $H(x) = \sqrt{x}$. Now, people with high conviction is more resistant to change (randomly) with respect to the previous case. On the other hand, if the conviction is small, $x < 1$, the individual is less resistant to change. Direct computations now show that the steady profile is given by

$$H_\infty(x) = H_0 x^{-1+(2\lambda_B M)/\mu} \exp\left\{-\frac{2\lambda}{\mu}x\right\}, \quad (74)$$

where the constant H_0 is chosen to fix the total mass of $H_\infty(x)$ equal to one. At difference with the previous case, the distribution decays exponentially to infinity, thus describing a population in which there are very few agents with a large conviction. Moreover, this distribution describes a population with a huge number of undecided agents. Note that, since the exponent of x in $H_\infty(\cdot)$ is strictly bigger than -1 , $H_\infty(\cdot)$ is integrable for any choice of the relevant parameters.

3.1.2 The Boltzmann Equation for Opinion and Conviction

In its original formulation (4) both the compromise and the self-thinking intensities were assigned in terms of the universal constant η and of the universal random parameters ξ, ξ_* . Suppose now that these quantities in (1) could depend of the personal conviction of the agent. For example, one reasonable assumption would be that an individual with high personal conviction is more resistant to move towards opinion of any other agent by compromise. Also, an high conviction could imply a reduction of the personal self-thinking. If one agrees with these assumptions, the binary trade (1) has to be modified to include the effect of conviction. Given two agents A and B characterized by the pair (x, w) (respectively (y, w_*)) of conviction and opinion, the new binary trade between A and B now reads

$$\begin{aligned} w' &= w - \eta \Psi(x) P(w)(w - w_*) + \Phi(x)\xi D(w), \\ w'_* &= w_* - \eta \Psi(y) P(w_*)(w_* - w) + \Phi(y)\xi_* D(w_*). \end{aligned} \quad (75)$$

In (75) the personal compromise propensity and self-thinking of the agents are modified by means of the functions $\Psi = \Psi(x)$ and $\Phi = \Phi(x)$, which depend on the convictions parameters. In this way, the outcome of the interaction results from a combined effect of (personal) compromise propensity, conviction and opinion. Among other possibilities, one reasonable choice is to fix the functions $\Psi(\cdot)$ and $\Phi(\cdot)$ as non-increasing functions. This reflects the idea that the conviction acts to increase the tendency to remain of the same opinion. Among others, a possible choice is

$$\Psi(x) = (1 + (x - A)_+)^{-\alpha}, \quad \Phi(x) = (1 + (x - B)_+)^{-\beta}.$$

Here A, B, α, β are nonnegative constants, and $h(x)_+$ denotes the positive part of $h(x)$. By choosing $A > 0$ (respectively $B > 0$), conviction will start to influence the change of opinion only when $x > A$ (respectively $x > B$). It is interesting to remark that the presence of the conviction parameter (through the functions Ψ and Φ), is such that the post-interaction opinion of an agent with high conviction remains close to the pre-interaction opinion. This induces a mechanism in which the opinions of agents with low conviction are attracted towards opinions of agents with high conviction.

Assuming the binary trade (75) as the microscopic binary exchange of conviction and opinion in the system of agents, the joint evolution of these quantities is described in terms of the density $f = f(x, w, t)$ of agents which at time $t > 0$ are represented by their conviction $x \in \mathbb{R}_+$ and wealth $w \in I$. The evolution in time of the density f is described by the following kinetic equation (in weak form) [89]

$$\begin{aligned} \frac{d}{dt} \int_{\mathbb{R}_+ \times I} \varphi(x, w) f(x, w, t) dx dw = \\ \frac{1}{2} \left(\int_{\mathbb{R}_+^2 \times I^2} (\varphi(x', w') + \varphi(y', w'_*) - \varphi(x, w) - \varphi(y, v)) \right. \\ \left. f(x, w, t) f(y, w_*, t) C(z) dx dy dz dw dw_* \right). \end{aligned} \quad (76)$$

In (76) the pairs (x', w) and (y', w_*) are obtained from the pairs (x, w) and (y, w_*) by (70) and (75). Note that, by choosing φ independent of w , that is $\varphi = \varphi(x)$, equation (76) reduces to the equation (72) for the marginal density of conviction $F(x, t)$.

To obtain analytic solutions to the Boltzmann-like equation (76) is prohibitive. The main reason is that the unknown density in the kinetic equation depends on two variables with different laws of interaction. In addition, while the interaction for conviction does not depend on the opinion variable, the law of interaction for the opinion does depend on the conviction. Also, at difference with the one-dimensional models, passage to Fokker-Planck equations (cf. [90] and the references therein) does not help in a substantial way. For this reason, we will resort to numerical investigation of (76), to understand the effects of the introduction of the conviction variable in the distribution of opinions.

3.1.3 Numerical Experiments

This section contains a numerical description of the solutions to the Boltzmann-type equation (76). For the numerical approximation of the Boltzmann equation we apply a Monte Carlo method, as described in Chapter 4 of [89]. If not otherwise stated the kinetic simulation has been performed with $N = 10^4$ particles.

The numerical experiments will help to clarify the role of conviction in the final distribution of the opinion density among the agents. The numerical simulations enhance the fact that the density $f(x, w, t)$ will rapidly converge towards a stationary

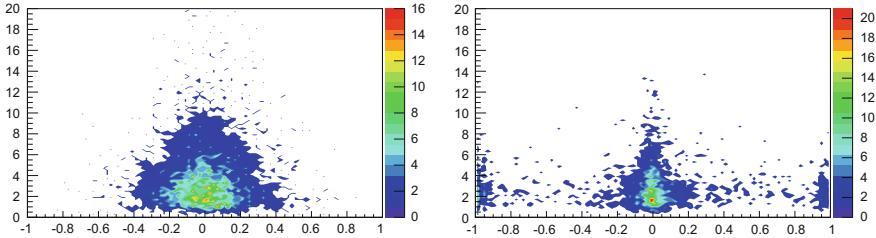


Fig. 6 Test 1: The particles solution with $N = 10000$ particles and linear H . High diffusion in conviction and reduced self-thinking (up) compared to low diffusion in conviction and high self-thinking (down)

distribution [89]. As usual in kinetic theory, this stationary solution will be reached in an exponentially fast time.

The numerical experiments will report the joint density of conviction and opinion in the agent system. The opinion variable will be reported on the horizontal axis, while the conviction variable will be reported on the vertical one. The color intensity will refer to the concentration of opinions. The following numerical tests have been considered.

Test 1

In the first test we consider the case of a conviction interaction where the diffusion coefficient in (70) is linear, $H(x) = x$. As described in Section 3.1.1 the distribution of conviction in this case is heavy tailed, with an important presence of agents with high conviction, and a large part of the population with a mean degree of conviction. In (75) we shall consider

$$\Phi(x) = \Psi(x) = \frac{1}{1 + (x - 1)_+}.$$

We further take $\lambda = \lambda_B = 0.5$ in (70), and $P(w) = 1$, $D(w) = \sqrt{1 - w^2}$ in (75). We consider a population of agents with an initially uniformly distributed opinion and a conviction uniformly distributed on the interval $[0, 5]$. We choose a time step of $\Delta t = 1$ and a final computation time of $t = 50$, where the steady state is practically reached.

Since the evolution of the conviction in the model is independent from the opinion, the latter is scaled in order to fix the mean equal to 0. We report the results for the particle density corresponding to different values of μ , η and the variance ξ^2 of the random variables ξ and ξ_* in Figure 6. This allows to verify the essential role of the diffusion processes in conviction and opinion formation.

Test 2

In this new test, we maintain the same values for the parameters, and we modify the diffusion coefficient in (70), which is now assumed as $H(x) = \sqrt{x}$. Within this

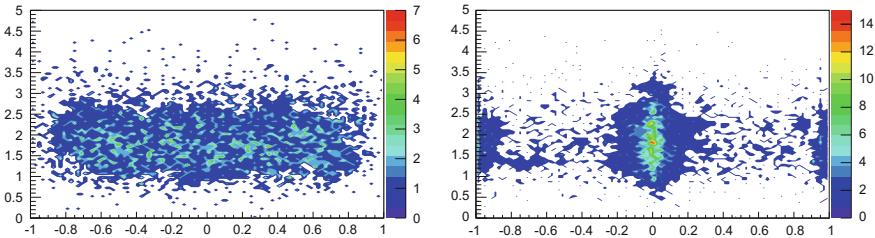


Fig. 7 Test 2: The particles solution with $N = 10000$ particles and $H(x) = \sqrt{x}$. High diffusion in conviction and reduced self-thinking (up) compared to low diffusion in conviction and high self-thinking (down).

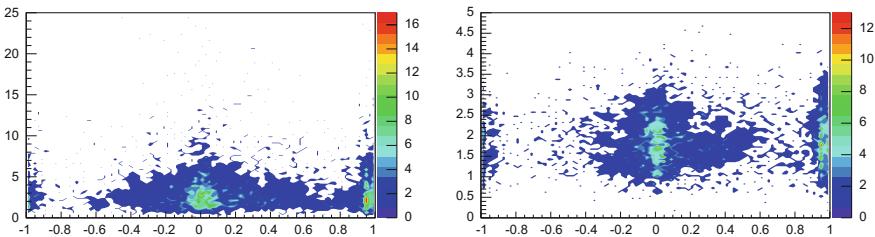


Fig. 8 Initial asymmetry in opinion leads to different opinion-conviction distributions. $H(x) = x$ (up), and $H(x) = \sqrt{x}$ (down).

choice, with respect to the previous test we expect the formation of a larger class on undecided agents. The results are reported in Figure 7 for the full density. At difference with the results of Test 1, opinion is spread out almost uniformly among people with low conviction. It is remarkable that in this second test, as expected, conviction is essentially distributed in the interval $[0, 5]$, at difference with Test 1, where agents reach a conviction parameter of 20.

The same effect is evident in Figure 8, which refers to both Tests 1 and 2 in which, to understand the evolution in case of asymmetry, the initial distribution of opinions was chosen uniformly distributed on the positive part of the interval.

3.2 Modeling Complex Networks

The present setting takes into account large complex networks of interacting agents by introducing a kinetic model which couples an alignment dynamics with the underlying evolution of the network. The coupled evolution of opinions and network is described by a Boltzmann-type equation where the probability distribution of opinions depends on a second relevant variable called *connectivity*. In principle, the ideas proposed here are not limited to a particular kind of opinion dynamics and one can easily adapt to the same situation other models developed in the literature [30, 59, 97].

3.2.1 A Boltzmann-Type Model for Opinion and Number of Connections

Let us consider a large system of agents interacting through a given network. We associate to each agent an opinion w , which varies continuously in $I = [-1, 1]$, and his number of connections c , a discrete variable varying between 0 and the maximum number of connections allowed by the network. Note that this maximum number typically is a fixed value which is several orders of magnitude smaller than the size the network.

We are interested in the evolution of the density function

$$f = f(w, c, t), \quad f : I \times \mathcal{C} \times \mathbb{R}^+ \rightarrow \mathbb{R}^+ \quad (77)$$

where $w \in I$ is the opinion variable, $c \in \mathcal{C} = \{0, 1, 2, \dots, c_{\max}\}$ is the discrete variable describing the number of connections and $t \in \mathbb{R}^+$ denotes as usual the time variable. For each time $t \geq 0$ the marginal density

$$\rho(c, t) = \int_I f(w, c, t) dw, \quad (78)$$

defines the evolution of the number of connections of the agents or equivalently the degree distribution of the network. In the sequel we assume that the total number of agents is conserved, i.e. $\sum_{c=0}^{c_{\max}} \rho(c, t) = 1$. The overall opinion distribution is defined likewise as the marginal density function

$$g(w, t) = \sum_{c=0}^{c_{\max}} f(w, c, t). \quad (79)$$

We express the evolution of the opinions by a binary interaction rule. From a microscopic point of view we suppose that the agents modify their opinion through binary interactions which depend on opinions and number of connections. If two agents with opinion and number of connections (w, c) and (w_*, c_*) meet, their post-interaction opinions are given by

$$\begin{cases} w' = w - \eta P(w, w_*; c, c_*)(w - w_*) + \xi D(w, c), \\ w'_* = w_* - \eta P(w_*, w; c_*, c)(w_* - w) + \xi_* D(w_*, c_*), \end{cases} \quad (80)$$

Note that, in the present setting the compromise function P depends both on the opinions and on the number of connections of each agent. In (80) all the other quantities are defined as in (1). We will consider by now a general interaction potential such that $0 \leq P(w, w_*, c, c_*) \leq 1$. In absence of diffusion $\xi, \xi_* \equiv 0$, and from (80) we have

$$|w' - w'_*| = |1 - \eta(P(w, w_*; c, c_*) + P(w_*, w; c_*, c))||w - w_*|. \quad (81)$$

Hence the post-exchange distances between agents are diminishing if we consider $\eta \in (0, 1)$ and $0 \leq P(w, w_*, c, c_*) \leq 1$. Similarly to Section 2.1, Proposition 1, we can require the conditions on the noise term to ensure that the post-interaction opinions do not leave the reference interval interval.

The evolution in time of the density function $f(w, c, t)$ is described by the following integro-differential equation of Boltzmann-type

$$\frac{d}{dt} f(w, c, t) + \mathcal{N}[f(w, c, t)] = Q(f, f)(w, c, t), \quad (82)$$

where $\mathcal{N}[\cdot]$ is an operator which is related to the evolution of the connections in the network and $Q(\cdot, \cdot)$ is the binary interaction operator. It is convenient to define Q in weak form as follows

$$\int_I Q(f, f) \varphi(w) dw = \lambda \sum_{c_*=0}^{c_{\max}} \left(\int_{I^2} (\varphi(w') - \varphi(w)) f(w_*, c_*) f(w, c) dw dw_* \right). \quad (83)$$

Consequently the equation (82) in weak form reads

$$\begin{aligned} \frac{d}{dt} \int_I f(w, c) \varphi(w) dw + \int_I \mathcal{N}[f(w, c)] \varphi(w) dw = \\ \frac{\lambda}{2} \sum_{c_*=0}^{c_{\max}} \left(\int_{I^2} (\varphi(w') + \varphi(w_*) - \varphi(w) - \varphi(w_*)) f(w_*, c_*) f(w, c) dw dw_* \right). \end{aligned} \quad (84)$$

3.2.2 Evolution of the Network

The operator $\mathcal{N}[\cdot]$ is defined through a combination of preferential attachment and uniform processes describing the evolution of the connections of the agents by removing and adding links in the network. These processes are strictly related to the generation of stationary scale-free distributions [14, 25, 105]. More precisely, for each $c = 1, \dots, c_{\max} - 1$ we define

$$\begin{aligned} \mathcal{N}[f(w, c, t)] = & - \frac{2V_r(f; w)}{\gamma + \beta} [(c + 1 + \beta) f(w, c + 1, t) - (c + \beta) f(w, c, t)] \\ & - \frac{2V_a(f; w)}{\gamma + \alpha} [(c - 1 + \alpha) f(w, c - 1, t) - (c + \alpha) f(w, c, t)], \end{aligned} \quad (85)$$

where $\gamma = \gamma(t)$ is the mean density of connectivity defined as

$$\gamma(t) = \sum_{c=0}^{c_{\max}} c \rho(c, t), \quad (86)$$

$\alpha, \beta > 0$ are attraction coefficients, and $V_r(f; w) \geq 0, V_a(f; w) \geq 0$ are characteristic rates of the removal and adding steps, respectively. The first term in (85) describes the net gain of $f(w, c, t)$ due to the connection removal between agents whereas the second term represents the net gain due to the connection adding process. The factor 2 has been kept in evidence since connections are removed and created pairwise. At the boundary we have the following equations

$$\begin{aligned}\mathcal{N}[f(w, 0, t)] &= -\frac{2V_r(f; w)}{\gamma + \beta}(\beta + 1)f(w, 1, t) + \frac{2V_a(f; w)}{\gamma + \alpha}\alpha f(w, 0, t), \\ \mathcal{N}[f(w, c_{\max}, t)] &= \frac{2V_r(f; w)}{\gamma + \beta}(c_{\max} + \beta)f(w, c_{\max}, t) \\ &\quad - \frac{2V_a(f; w)}{\gamma + \alpha}(c_{\max} - 1 + \alpha)f(w, c_{\max} - 1, t),\end{aligned}\tag{87}$$

which are derived from (85) taking into account the fact that, in the dynamics of the network, connections cannot be removed from agents with 0 connections and cannot be added to agents with c_{\max} connections.

The evolution of the connections of the network can be recovered taking $\varphi(w) = 1$ in (84)

$$\frac{d}{dt}\rho(c, t) + \int_I \mathcal{N}[f(w, c, t)] dw = 0.\tag{88}$$

From the definition of the network operator $\mathcal{N}[\cdot]$ it follows that

$$\frac{d}{dt} \sum_{c=0}^{c_{\max}} \rho(c, t) = 0.\tag{89}$$

Then, with the collisional operator defined in (83) and of $\mathcal{N}[\cdot]$ in (85) the total number of agents is conserved.

Let us take into account the evolution of the mean density of connectivity γ defined in (86). For each $t \geq 0$

$$\begin{aligned}\frac{d}{dt}\gamma(t) &= -2 \int_I V_r(f; w) \frac{\gamma_f + \beta g(w, t)}{\gamma + \beta} dw + 2 \int_I V_a(f; w) \frac{\gamma_f + \alpha g(w, t)}{\gamma + \alpha} dw \\ &\quad + \frac{2\beta}{\gamma + \beta} \int_I V_r(f; w) f(w, 0, t) dw - \frac{2(c_{\max} + \alpha)}{\gamma + \alpha} \int_I V_a(f; w) f(w, c_{\max}, t) dw.\end{aligned}\tag{90}$$

Therefore $\gamma(t)$ is in general not conserved. The explicit computations for the conservation of the total number of connections and for the evolution of the mean density of connectivity are reported under specific assumptions in [10].

When V_a and V_r are constants, the operator $\mathcal{N}[\cdot]$ is linear and will be denoted by $\mathcal{L}[\cdot]$. In this case, the evolution of the network of connections is independent from the opinion and one gets the closed form

$$\frac{d}{dt}\rho(c, t) + \mathcal{L}[\rho(c, t)] = 0, \quad (91)$$

where

$$\begin{aligned} \mathcal{L}[\rho(c, t)] = & -\frac{2V_r}{\gamma + \beta} [(c + 1 + \beta)\rho(c + 1, t) - (c + \beta)\rho(c, t)] \\ & - \frac{2V_a}{\gamma + \alpha} [(c - 1 + \alpha)\rho(c - 1, t) - (c + \alpha)\rho(c, t)], \end{aligned} \quad (92)$$

with the boundary conditions

$$\begin{aligned} \mathcal{L}[\rho(0, t)] = & -\frac{2V_r}{\gamma + \beta}(\beta + 1)\rho(1, t) + \frac{2V_a}{\gamma + \alpha}\alpha\rho(0, t), \\ \mathcal{L}[\rho(c_{\max}, t)] = & \frac{2V_r}{\gamma + \beta}(c_{\max} + \beta)\rho(c_{\max}, t) - \frac{2V_a}{\gamma + \alpha}(c_{\max} - 1 + \alpha)\rho(c_{\max} - 1, t). \end{aligned} \quad (93)$$

Note that in (92) the dynamics correspond to a combination of preferential attachment processes ($\alpha, \beta \approx 0$) and uniform processes ($\alpha, \beta \gg 1$) with respect to the probability density of connections $\rho(c, t)$. Concerning the large time behavior of the network of connections, in the linear case with $V_r = V_a$, $\beta = 0$ and denoting by γ the asymptotic value of the density of connectivity it holds

Proposition 2 *For each $c \in \mathcal{C}$ the stationary solution to (91) or equivalently*

$$(c + 1)\rho_\infty(c + 1) = \frac{1}{\gamma + \alpha} [(c(2\gamma + \alpha) + \gamma\alpha)\rho_\infty(c) - \gamma(c - 1 + \alpha)\rho_\infty(c - 1)], \quad (94)$$

is given by

$$\rho_\infty(c) = \left(\frac{\gamma}{\gamma + \alpha}\right)^c \frac{1}{c!} \alpha(\alpha + 1) \cdots (\alpha + c - 1) \rho_\infty(0), \quad (95)$$

where

$$\rho_\infty(0) = \left(\frac{\alpha}{\alpha + \gamma}\right)^\alpha. \quad (96)$$

A detailed proof is given in [10]. Further approximations are possible if $\alpha \gg 1$ or $\alpha \approx 0$. For large α the preferential attachment process described by the master equation (92) is destroyed and the network approaches a random network, whose degree distribution coincides with the Poisson distribution. In fact, in the limit $\alpha \rightarrow +\infty$ we have $(\alpha + \gamma)^c \approx \alpha(\alpha + 1) \cdots (\alpha + c - 1)$, and

$$\rho_\infty(c) = \lim_{\alpha \rightarrow +\infty} \left(1 + \frac{\gamma}{\alpha}\right)^{-\alpha} \gamma^c = \frac{e^{-c}}{c!} \gamma^c. \quad (97)$$

In the second case, for $\gamma \geq 1$ and small values of α , the distribution can be correctly approximated with a truncated power-law with unitary exponent

$$\rho_\infty(c) = \left(\frac{\alpha}{\gamma}\right)^\alpha \frac{\alpha}{c!}. \quad (98)$$

3.2.3 Fokker-Planck Modeling

Similarly to Section 2.1 we can derive a Fokker-Planck equation through the quasi-invariant opinion limit. Let us introduce the scaling parameter $\varepsilon > 0$ and consider the scaling

$$\eta = \varepsilon, \quad \lambda = \frac{1}{\varepsilon}, \quad \varsigma^2 = \varepsilon \sigma^2. \quad (99)$$

In the limit $\varepsilon \rightarrow 0$ we obtain the Fokker-Planck equation for the evolution of the opinions' distribution through the evolving network

$$\frac{\partial}{\partial t} f(w, c, t) + \mathcal{N}[f(w, c, t)] = \frac{\partial}{\partial w} \mathcal{P}[f] f(w, c, t) + \frac{\sigma^2}{2} \frac{\partial^2}{\partial w^2} (D(w, c)^2 f(w, c, t)), \quad (100)$$

where

$$\mathcal{P}[f](w, c, t) = \sum_{c_*=0}^{c_{\max}} \int_I P(w, w_*; c, c_*) (w_* - w) f(w_*, c_*, t) dw_*. \quad (101)$$

In some case it is possible to compute explicitly the steady state solutions of the Fokker-Planck system (100). We restrict to linear operators $\mathcal{L}[\cdot]$ and asymptotic solutions of the following form

$$f_\infty(w, c) = g_\infty(w) \rho_\infty(c), \quad (102)$$

where $\rho_\infty(c)$ is the steady state distribution of the connections (see Proposition 2) and

$$\int_I f_\infty(w, c) dw = \rho_\infty(c), \quad \sum_{c=0}^{c_{\max}} f_\infty(w, c) = g_\infty(w). \quad (103)$$

From the definition of the linear operator $\mathcal{L}[\cdot]$ we have $\mathcal{L}[\rho_\infty(c)] = 0$. Hence the stationary solutions of type (102) satisfy the equation

$$\frac{\partial}{\partial w} \mathcal{P}[f_\infty] f_\infty(w, c) + \frac{\sigma^2}{2} \frac{\partial^2}{\partial w^2} (D(w, c)^2 f_\infty(w, c)) = 0. \quad (104)$$

Equation (104) can be solved explicitly in some case [7, 98]. If P is in the form

$$P(w, w_*; c, c_*) = H(w, w_*) K(c, c_*), \quad (105)$$

the operator $\mathcal{P}[f_\infty]$ can be written as follows

$$\mathcal{P}[f_\infty](w, c) = \left(\sum_{c_*=0}^{c_{\max}} K(c, c_*) \rho_\infty(c_*) \right) \left(\int_I H(w, w_*) (w_* - w) g_\infty(w_*) dw_* \right). \quad (106)$$

We further assume that $K(c, c_*) = K(c_*)$ is independent of c and denote

$$\kappa = \sum_{c_*=0}^{c_{\max}} K(c_*) \rho_\infty(c_*), \quad \bar{m}_w = \sum_{c=0}^{c_{\max}} m_w(c, t), \quad m_w(c, t) = \int_I w f(w, c, t) dw.$$

a) In the case $H \equiv 1$ and $D(w) = 1 - w^2$ the steady state solution g_∞ is given by

$$g_\infty(w) = C_0 (1 + w)^{-2 + \bar{m}_w \kappa / \sigma^2} (1 - w)^{-2 - m_w \kappa / \sigma^2} \exp \left\{ - \frac{\kappa (1 - \bar{m}_w w)}{\sigma^2 (1 - w^2)} \right\}, \quad (107)$$

b) For $H(w, w_*) = 1 - w^2$ and $D(w) = 1 - w^2$ the steady state solution g_∞ is given by

$$g_\infty(w) = C_0 (1 - w)^{-2 + (1 - \bar{m}_w) \kappa / \sigma^2} (1 + w)^{-2 + (1 + \bar{m}_w) \kappa / \sigma^2}, \quad (108)$$

In Figure 9 we report the stationary solution $f_\infty(w, c) = g_\infty(w) \rho_\infty(c)$, where $g_\infty(w)$ is given by (107) with $\kappa = 1$, $m_w = 0$, $\sigma^2 = 0.05$ and $p_\infty(c)$ defined by (95), with $V_r = V_a = 1$, $\gamma = 30$ and $\alpha = 10$ on the left and $\alpha = 0.01$ on the right.

3.2.4 Numerical Experiments

In this section we perform some numerical experiments to study the behavior of the new kinetic models. We focus on the case $\alpha < 1$, since it represents the most relevant case in complex networks [2, 105]. Within this range of the parameter we have emergence of power law distributions for network's connectivity. In the tests that follow the opinion dynamics evolves according to (100). The compromise function P and the local diffusion function D in the various tests will be specified in each test. The different tests are summarized in Table 1, where other parameters are introduced and additional details are reported. In Test 1 a Monte Carlo method is used to solve the Boltzmann model (82) we refer to [5, 88, 89] and to the Appendix for a description on these class of methods. In Test 2, 3, 4 the Fokker-Planck system (100) is solved

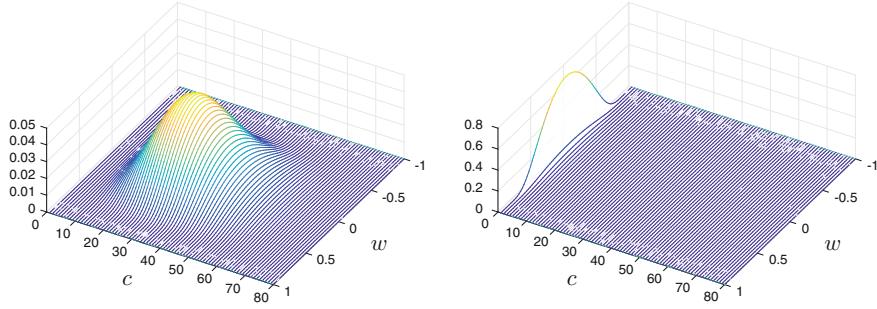


Fig. 9 Stationary solutions of type $f_{\infty}(w, c) = g_{\infty}(w)p_{\infty}(c)$, where $g_{\infty}(w)$ is given by (107) with $\kappa = 1$, $m_w = 0$, $\sigma^2 = 0.05$ and $p_{\infty}(c)$ defined by (95), with $V_r = V_a = 1$, and $\alpha = 10$ on the left and $\alpha = 0.1$ on the right.

Table 2 Parameters in the various test cases

Test	σ^2	σ_F^2	σ_L^2	c_{\max}	V_r	V_a	γ_0	α	β
#1	5×10^{-2}	6×10^{-2}	—	250	1	1	30	1×10^{-1}	0
#2	5×10^{-3}	4×10^{-2}	2.5×10^{-2}	250	1	1	30	1×10^{-4}	0
#3	1×10^{-3}	—	—	250	1	1	30	1×10^{-1}	0

via the steady-state preserving Chang-Cooper scheme, see the Appendix and [32, 33, 39, 76, 83] for further details.

Test 1

We first consider the one dimensional setting to show the convergence of the Boltzmann model (82) to the exact solution of the Fokker-Planck system (100), via Monte Carlo methods. We simulate the dynamics with the linear interaction kernel, $P(w, w_*; c, c_*) = 1$, and $D(w, c) = 1 - w^2$, thus we can use the results (107), to compare the solutions obtained through the numerical scheme with the analytical one, the other parameters of the model are reported in Table 2 and we define the following initial data

$$g_0(w) = \frac{1}{2\sqrt{2\pi\sigma_F^2}}(\exp\{-(w + 1/2)^2/(2\sigma_F^2)\} + \exp\{-(w - 1/2)^2/(2\sigma_F^2)\}). \quad (109)$$

In Figure 10, on the left hand-side, we report the qualitative convergence of the Binary Interaction algorithm, [5], where we consider $N_s = 10^5$ samples to reconstruct the opinion's density, $g(w, t)$, on a grid of $N = 80$ points. The figure shows that for decreasing values of the scaling parameter $\varepsilon = \{0.5, 0.05, 0.005\}$, we have convergence to the reference solutions, (107) of the Fokker-Planck equation. On the right we report the convergence to the stationary solution of the connectivity distribution, (94), for $\alpha = 0.1$ and $V = 1$ and with $c_{\max} = 250$. In this case we show two

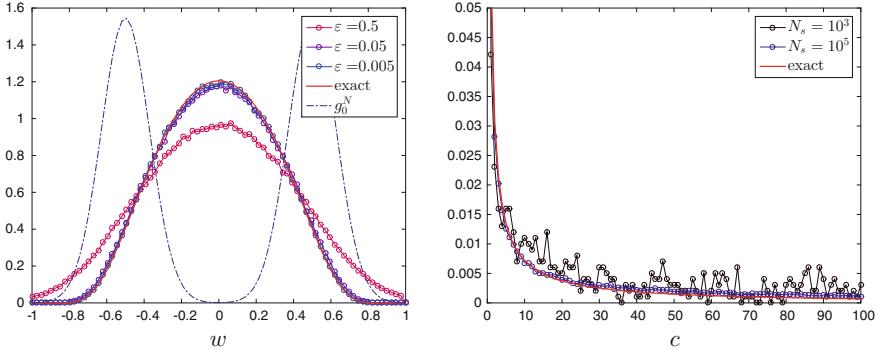


Fig. 10 Test 1. One-dimensional setting: on the left, convergence of (82) to the stationary solution (107), of the Fokker-Planck equation, for decreasing values of the parameter ϵ , g_0^N represent the initial distribution. On the right, convergence of the Monte-Carlo Algorithm 1, see the Appendix, to the reference solution (94) for increasing value of the the number of samples N_s .

different qualitative behaviors for an increasing number of samples $N_s = \{10^3, 10^5\}$ and for sufficient large times, obtained through the stochastic Algorithm 1.

Test 2

In the second test we analyze the influence of the connections over the opinion dynamics, for a compromise function of the type (105) where $H(w, w_*) = 1 - w^2$ and K defined by

$$K(c, c_*) = \left(\frac{c}{c_{\max}} \right)^{-a} \left(\frac{c_*}{c_{\max}} \right)^b, \quad (110)$$

for $a, b > 0$. This type of kernel assigns higher relevance into the opinion dynamics to higher connectivity, and low influence to low connectivity. The diffusivity is weighted by $D(w, c) = 1 - w^2$. We perform a first computation with the initial condition

$$f_0(w, c) = C_0 \begin{cases} \rho_\infty(c) \exp\{-(w + \frac{1}{2})^2/(2\sigma_F^2)\}, & \text{if } 0 \leq c \leq 20, \\ \rho_\infty(c) \exp\{-(w - \frac{3}{4})^2/(2\sigma_L^2)\}, & \text{if } 60 \leq c \leq 80, \\ 0, & \text{otherwise.} \end{cases} \quad (111)$$

The values of the parameters are reported in the third line of Table 1. In the interaction function $K(\cdot, \cdot)$ in (110) we choose $a = b = 3$. The evolution is performed through the Chang-Cooper type scheme with $\Delta w = 2/N$ and $N = 80$. The evolution of the system is studied in the time interval $[0, T]$, with $T = 2.5$.

In Figure 11 we report the result of the simulation. On the first plot the initial configuration is split in two parts, the majority concentrated around the opinion $\bar{w}_F = -1/2$ and only a small portion concentrated around $\bar{w}_L = 3/4$. We observe that, because of the anisotropy induced by $K(c, c_*)$, the density with a low level of

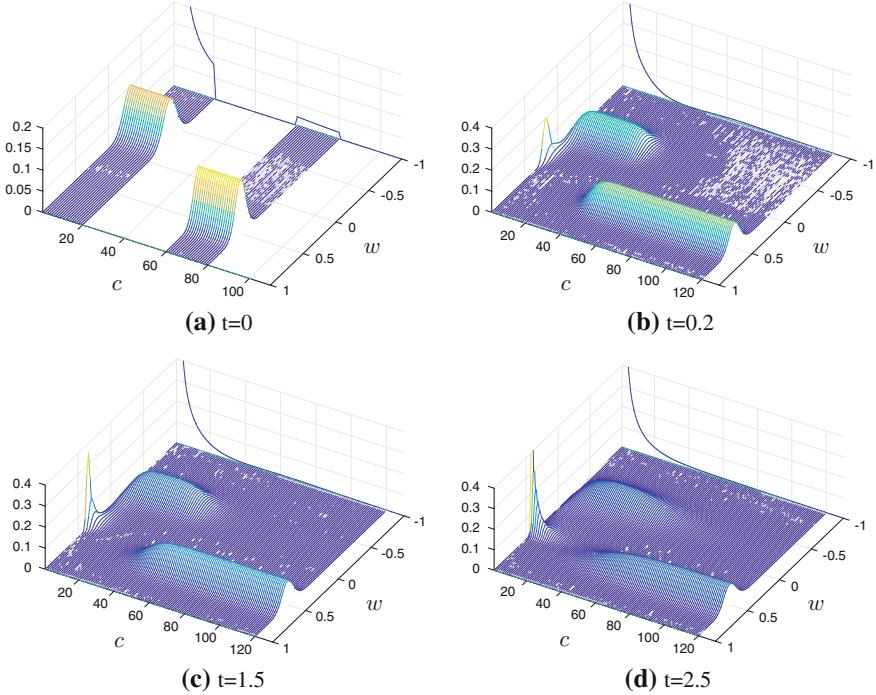


Fig. 11 Test #2. Evolution of the initial data (111) in the time interval $[0, T]$, with $T = 2.5$. The evolution shows how a small portion of density with high connectivity can bias the majority of the population towards their position. (Note: The density is scaled according to the marginal distribution $\rho(c, t)$ in order to better show its evolution. The actual marginal density $\rho(c, t)$ is depicted in the background, scaled by a factor 10).

connectivity is influenced by the small concentration of density around w_L with a large level of connectivity.

Test 3

Finally, we consider the Hegselmann-Krause model, [72], known also as bounded confidence model, where agents interact only with agents whose opinion lays within a certain range of confidence. Thus we define the following compromise function

$$P(w, w_*; c, c_*) = \chi_{\{|w-w_*| \leq \Delta(c)\}}(w_*) \quad \text{with} \quad \Delta(c) = d_0 \frac{c}{c_{\max}},$$

where the confidence level, $\Delta(c)$, is assumed to depend on the number of connections, so that agents with higher number of connections are prone to larger level of confidence. We define the initial data

$$f_0(w, c) = \frac{1}{2} \rho_\infty(c), \quad (112)$$

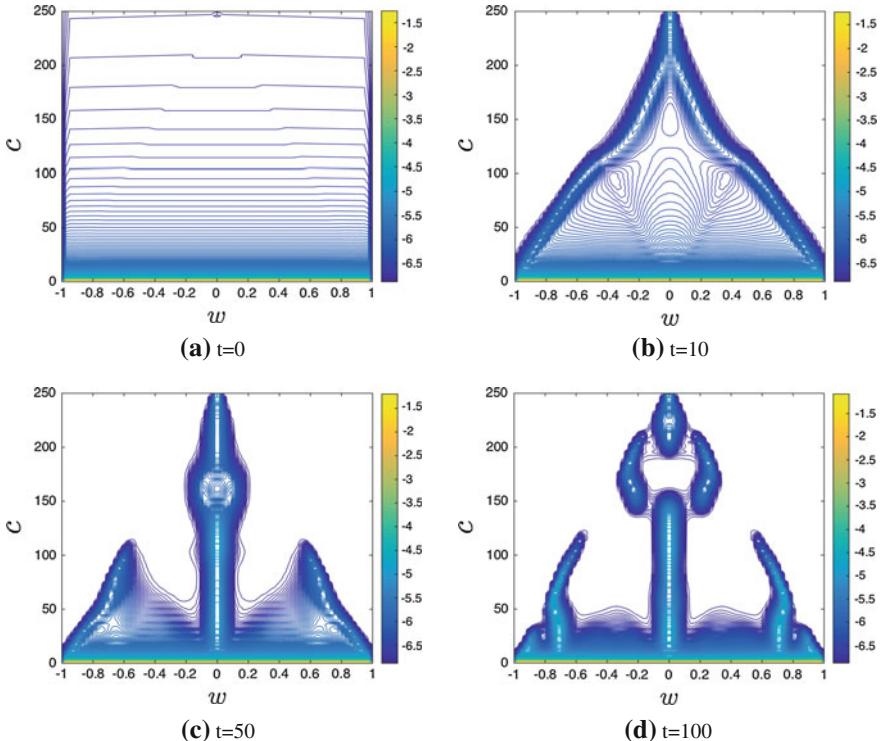


Fig. 12 Test #3. Evolution of the solution of the Fokker-Planck model (100), for the bounded confidence model, with $\Delta(c) = d_0 c / c_{\max}$, and $d_0 = 1.01$, in the time frame $[0, T]$, with $T = 100$. The choice of $\Delta(c)$ reflects in the heterogeneous emergence of clusters with respect to the connectivity level: for higher level of connectivity consensus is reached, instead for lower levels of connectivity multiple opinion clusters are present. (Note: In order to better show its evolution, we represent the solution as $\log(f(w, c, t) + \varepsilon)$, with $\varepsilon = 0.001$.)

therefore the opinion is uniformly distributed on the interval $I = [-1, 1]$ and it decreases along $c \in [0, c_{\max}]$ following $\rho_\infty(c)$, as in (94), with parameters defined in Table 2 and $D(w, c) = 1 - w^2$. Figure 12 shows the evolution of (112), where $\Delta(c)$ creates an heterogeneous emergence of clusters with respect to the connectivity level: for higher level of connectivity consensus is reached, since the bounded confidence level is larger, instead for lower levels of connectivity multiple clusters appears, up to the limiting case $c = 0$, where the opinions are not influenced by the consensus dynamics.

4 Final Considerations

The mathematical modeling of opinion formation in multi-agent systems is nowadays a well studied field of application of kinetic theory. Starting from some basic models [89, 98] we tried to enlighten some recent improvements, in which the models have been enriched by adding further aspects with the goal to better reflect various facts of our daily life. Particular emphasis has been done to control strategies on opinion formation, a theme of paramount importance with several potential applications. Testimonials in advertising a product or opinion leaders during elections may lead the group of interacting agents towards a desired state and practically can modify the way a society behaves and is ruled by a government. Furthermore, conviction has been shown to be important in order to achieve a final personal opinion, in view of the fact that a society with a high number of stubborn people clearly behaves very differently from a society composed by very susceptible persons. Last, the recent development and increasing importance of social networks made the study of opinions in this area a crucial and actual research theme.

Finally, we mention here that the idea to study the role of opinion leaders in this kinetic setting has been first studied in [59] and subsequently improved by considering another independent variable which quantifies leadership qualities in [58]. Clearly, our point of view and the material we presented here gives only a selected partial view of the whole research in the field. The interested reader can however find an almost exhaustive list of references which could certainly help to improve his knowledge on this interesting subject.

Acknowledgements This work has been written within the activities of the National Groups of Scientific Computing (GNCS) and Mathematical Physics (GNFM) of the National Institute of High Mathematics of Italy (INDAM). GA acknowledges the ERC-Starting Grant project High-Dimensional Sparse Optimal Control (HDSPCONTR). GT acknowledges the partial support of the MIUR project *Optimal mass transportation, geometrical and functional inequalities with applications*.

Appendix: Numerical Simulation Methods

In this short appendix we sketch briefly some particular numerical technique used to produce the various simulation results presented in the manuscript. We omit the description of the Monte Carlo simulation approach for the Boltzmann equation describing the opinion exchange dynamics addressing the interested reader to [89]. For the development of Monte Carlo methods that works in the Fokker-Planck regime we refer to [5].

We first summarize the Monte Carlo approach used to deal with the evolution of the social network and then the steady state preserving finite-difference approach used for the mean-field models. More details can be found in [10].

Monte Carlo Algorithm for the Evolution of the Network

The evolution of the network is given by

$$\begin{cases} \frac{d}{dt} f(w, c, t) + \mathcal{N}[f(w, c, t)] = 0, \\ f(w, c, 0) = f_0(w, c). \end{cases}$$

Let $f^n = f(w, c, t^n)$ the empirical density function for the density of agents at time t^n with opinion w and connections c . For any given opinion w we approximate the solution of the above problem at time t^{n+1} by

$$\begin{aligned} f^{n+1}(w, c) &= \left(1 - \Delta t \frac{V_r(c + \beta)}{\gamma^n + \beta} - \Delta t \frac{V_a(c + \alpha)}{\gamma^n + \alpha} \right) f^n(w, c) \\ &\quad + \Delta t \frac{V_r(c + \beta)}{\gamma^n + \beta} f^n(w, c - 1) + \Delta t \frac{V_a(c + \alpha)}{\gamma^n + \alpha} f^n(w, c + 1), \end{aligned}$$

with boundary conditions

$$\begin{aligned} f^n(w, 0) &= \left(1 - \Delta t \frac{V_a(c + \alpha)}{\gamma^n + \alpha} \right) f^n(w, 0) + \Delta t \frac{V_a(c + \alpha)}{\gamma^n + \alpha} f^n(w, 1), \\ f^n(w, c_{\max}) &= \left(1 - \Delta t \frac{V_r(c + \beta)}{\gamma^n + \beta} \right) f^n(w, c_{\max}) + \Delta t \frac{V_r(c_{\max} + \beta)}{\gamma^n + \beta} f^n(w, c_{\max} - 1), \end{aligned}$$

and temporal discretization such that

$$\Delta t \leq \min \left\{ \frac{\gamma^n + \beta}{V_r(c_{\max} + \beta)}, \frac{\gamma^n + \alpha}{V_a(c_{\max} + \alpha)} \right\}. \quad (\text{A3})$$

The algorithm to simulate the above equation reads as follows

Algorithm 1

1. Sample (w_i^0, c_i^0) , with $i = 1, \dots, N_s$, from the distribution $f^0(w, c)$.
2. for $n = 0$ to $n_{tot} - 1$
 - a. Compute $\gamma^n = \frac{1}{N_s} \sum_{j=1}^{N_s} c_j^n$;
 - b. Fix Δt such that condition (A3) is satisfied.
 - c. for $k = 1$ to N_s
 - i. Compute the following probabilities rates

$$p_k^{(a)} = \frac{\Delta t V_a(c_k^n + \alpha)}{\gamma^n + \alpha}, \quad p_k^{(r)} = \frac{\Delta t V_r(c_k^n + \beta)}{\gamma^n + \beta},$$

- ii. Set $c_k^* = c_k^n$.

iii. if $0 \leq c_k^* \leq c_{max} - 1$,
 with probability $p_k^{(a)}$ add a connection: $c_k^* = c_k^* + 1$;
 iv. if $1 \leq c_k^* \leq c_{max}$,
 with probability $p_k^{(r)}$ remove a connection: $c_k^* = c_k^* - 1$;
 end for
 d. set $c_i^{n+1} = c_i^*$, for all $i = 1, \dots, N_s$.
 end for

Chang-Cooper Type Numerical Schemes

In the domain $(w, c) \in I \times \mathcal{C}$ we consider the Fokker-Planck system

$$\frac{\partial}{\partial t} f(w, c, t) + \mathcal{N}[f(w, c, t)] = \frac{\partial}{\partial w} \mathcal{F}[f], \quad (\text{A4})$$

with zero flux boundary condition on w , initial data $f(w, c, 0) = f_0(w, c)$ and

$$\mathcal{F}[f] = (\mathcal{P}[f] + \sigma^2 D'(w, c) D(w, c)) f(w, c, t) + \frac{\sigma^2}{2} D(w, c)^2 \frac{\partial}{\partial w} f(w, c, t),$$

where $\mathcal{P}[f]$ is given by (101). Let us introduce a uniform grid $w_i = -1 + i \Delta w$, $i = 0, \dots, N$ with $\Delta w = 2/N$, we denote by $w_{i \pm 1/2} = w_i \pm \Delta w/2$ and define

$$f_i(c, t) = \frac{1}{\Delta w} \int_{w_{i+1/2}}^{w_{i-1/2}} f(w, c, t) dw.$$

Integrating equation (A4) yields

$$\frac{\partial}{\partial t} f_i(c, t) + \mathcal{N}[f_i(c, t)] = \frac{\mathcal{F}_{i+1/2}[f] - \mathcal{F}_{i-1/2}[f]}{\Delta w},$$

where $\mathcal{F}_i[f]$ is the flux function characterizing the numerical discretization. We assume the Chang-Cooper flux function

$$\begin{aligned} \mathcal{F}_{i+1/2}[f] = & \left((1 - \delta_{i+1/2})(\mathcal{P}[f_{i+1/2}] + \sigma^2 D'_{i+1/2} D_{i+1/2}) + \frac{\sigma^2}{2 \Delta w} D_{i+1/2}^2 \right) f_{i+1} \\ & + \left(\delta_{i+1/2}(\mathcal{P}[f_{i+1/2}] + \sigma^2 D'_{i+1/2} D_{i+1/2}) - \frac{\sigma^2}{2 \Delta w} D_{i+1/2}^2 \right) f_i, \end{aligned}$$

where $D_{i+1/2} = D(w_{i+1/2}, c)$ and $D'_{i+1/2} = D'(w_{i+1/2}, c)$. The weights $\delta_{i+1/2}$ have to be chosen in such a way that a steady state solution is preserved. Moreover this

choice permits also to preserve nonnegativity of the numerical density. The choice

$$\delta_{i+1/2} = \frac{1}{\lambda_{i+1/2}} + \frac{1}{1 - \exp(\lambda_{i+1/2})}, \quad (\text{A5})$$

where

$$\lambda_{i+1/2} = \frac{2\Delta w}{\sigma^2} \frac{1}{D_{i+1/2}^2} (\mathcal{P}[f_{i+1/2}] + \sigma^2 D'_{i+1/2} D_{i+1/2}),$$

leads to a second order Chang-Cooper nonlinear approximation of the original problem. Note that here, at variance with the standard Chang-Cooper scheme [39], the weights depend on the solution itself as in [76]. Thus we have a nonlinear scheme which preserves the steady state with second order accuracy. In particular, by construction, the weight in (A5) are nonnegative functions with values in $[0, 1]$.

Higher order accuracy of the steady state can be recovered using a more general numerical flux [10].

References

1. D. Acemoglu, O. Asuman. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1, 3–49, 2011.
2. R. Albert, A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1): 1–47, 2002.
3. G. Albi, M. Bongini, E. Cristiani, D. Kalise. Invisible control of self-organizing agents leaving unknown environments. *SIAM Journal on Applied Mathematics*, to appear.
4. G. Albi, L. Pareschi. Modeling of self-organized systems interacting with a few individuals: from microscopic to macroscopic dynamics. *Applied Mathematics Letters*, 26: 397–401, 2013.
5. G. Albi, L. Pareschi. Binary interaction algorithm for the simulation of flocking and swarming dynamics. *SIAM Journal on Multiscale Modeling and Simulations*, 11(1), 1–29, 2013.
6. G. Albi, M. Herty, L. Pareschi. Kinetic description of optimal control problems and applications to opinion consensus. *Communications in Mathematical Sciences*, 13(6): 1407–1429, 2015.
7. G. Albi, L. Pareschi, M. Zanella. Boltzmann-type control of opinion consensus through leaders. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 372(2028): 20140138, 2014.
8. G. Albi, L. Pareschi, M. Zanella. Uncertainty quantification in control problems for flocking models. *Mathematical Problems in Engineering*, 2015, 14 pp., 2015.
9. G. Albi, L. Pareschi, M. Zanella. On the optimal control of opinion dynamics on evolving networks. *IFIP TC7 2015 Proceedings*, to appear Kinetic and Related Models, 10(1): 1–32, 2017.
10. G. Albi, L. Pareschi, M. Zanella. Opinion dynamics over complex networks: kinetic modeling and numerical methods. To appear in *Kinetic and related models*, 2016.
11. G. Aletti, G. Naldi, G. Toscani. First-order continuous models of opinion formation. *SIAM Journal on Applied Mathematics*, 67(3): 837–853, 2007.
12. L. A. N. Amaral, A. Scala, M. Bathémy, H.E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21): 11149–11152, 2000.
13. D. Armbruster, C. Ringhofer. Thermalized kinetic and fluid models for re-entrant supply chains. *Multiscale Modeling & Simulation*, 3(4): 782–800, 2005.

14. A.-L. Barabási, R. Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999.
15. A.-L. Barabási, R. Albert, H. Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1): 173–187, 1999.
16. N. Bellomo, G. Ajmone Marsan, A. Tosin. *Complex Systems and Society. Modeling and Simulation*. SpringerBriefs in Mathematics, Springer, 2013.
17. N. Bellomo, J. Soler. On the mathematical theory of the dynamics of swarms viewed as complex systems. *Mathematical Models and Methods in Applied Sciences*, 22(01): 1140006, 2012.
18. E. Ben-Naim. Opinion dynamics: rise and fall of political parties. *Europhysics Letters*, 69(5): 671, 2005.
19. E. Ben-Naim, P. L. Krapivski, S. Redner. Bifurcations and patterns in compromise processes. *Physica D: Nonlinear Phenomena*, 183(3): 190–204, 2003.
20. E. Ben-Naim, P. L. Krapivski, R. Vazquez, S. Redner. Unity and discord in opinion dynamics. *Physica A*, 330(1–2): 99–106, 2003.
21. A. Bensoussan, J. Frehse, P. Yam. Mean field games and mean field type control theory. *SpringerBriefs in Mathematics*, New York, NY: Springer, 2013.
22. M. L. Bertotti, M. Delitala. On a discrete generalized kinetic approach for modeling persuader's influence in opinion formation processes. *Mathematical and Computer Modeling*, 48(7–8): 1107–1121, 2008.
23. S. Biswas. Mean-field solutions of kinetic-exchange opinion models. *Physical Review E*, 84(5), 056105, 2011.
24. M. Bongini, M. Fornasier, F. Rossi, F. Solombrino. Mean-Field Pontryagin Maximum Principle, preprint, 2015.
25. C. M. Bordogna, E. V. Albano. Dynamic behavior of a social model for opinion formation. *Physical Review E*, 76(6): 061125, 2007.
26. A. Borzì, S. Wongkaew. Modeling and control through leadership of a refined flocking system. *Mathematical Models and Methods in Applied Sciences*, 25(2): 255–282, 2015.
27. L. Boudin, F. Salvarani. The quasi-invariant limit for a kinetic model of sociological collective behavior. *Kinetic and Related Models*: 433–449, 2009.
28. L. Boudin, F. Salvarani. A kinetic approach to the study of opinion formation. *ESAIM: Mathematical Modeling and Numerical Analysis*, 43(3): 507–522, 2009.
29. L. Boudin, F. Salvarani. Conciliatory and contradictory dynamics in opinion formation. *Physica A: Statistical Mechanics and its Applications*, 391(22): 5672–5684, 2012.
30. L. Boudin, R. Monaco, F. Salvarani. Kinetic model for multidimensional opinion formation. *Physical Review E*, 81(3): 036109, 2010.
31. C. Brugna, G. Toscani. Kinetic models of opinion formation in the presence of personal conviction. *Physical Review E*, 92, 052818, 2015.
32. C. Buet, S. Dellacherie. On the Chang and Cooper numerical scheme applied to a linear Fokker-Planck equation. *Communications in Mathematical Sciences*, 8(4): 1079–1090, 2010.
33. C. Buet, S. Cordier, V. Dos Santos. A conservative and entropy scheme for a simplified model of granular media. *Transport Theory and Statistical Physics*, 33(2): 125–155, 2004.
34. M. Burger, M. Di Francesco, P. A. Markowich, M.-T. Wolfram. Mean-field games with non-linear mobilities in pedestrian dynamics. *Discrete and Continuous Dynamical Systems - B*, 19(5): 1311–1333, 2014.
35. E. F. Camacho, C. Bordons. *Model Predictive Control*, Springer-Verlag London, 2004.
36. M. Caponigro, M. Fornasier, B. Piccoli, E. Trélat. Sparse stabilization and optimal control of the Cucker-Smale model. *Mathematical Control and Related Fields*, 3(4): 447–466, 2013.
37. C. Castellano, S. Fortunato, V. Loreto. Statistical physics of social dynamics. *Review of Modern Physics*, 81(2): 591–646, 2009.
38. A. Chakraborti, B. K. Chakrabarti. Statistical mechanics of money: how saving propensity affects its distribution. *European Physical Journal B*, 17: 167–170, 2000.
39. J. S. Chang, G. Cooper. A practical difference scheme for Fokker-Planck equation. *Journal of Computational Physics*, 6: 1–16, 1970.

40. H. Choi, M. Hinze, K. Kunisch. Instantaneous control of backward-facing step flows. *Applied Numerical Mathematics*, 31(2): 133–158, 1999.
41. R. M. Colombo, N. Pogodaev. Confinement strategies in a model for the interaction between individuals and a continuum. *SIAM Journal on Applied Dynamical Systems*, 11(2): 741–770, 2012.
42. R. M. Colombo, N. Pogodaev. On the control of moving sets: positive and negative confinement results. *SIAM Journal on Control and Optimization*, 51(1): 380–401, 2013.
43. V. Comincioli, L. Della Croce, G. Toscani. A Boltzmann-like equation for choice formation. *Kinetic and Related Models*, 2(1): 135–149, 2009.
44. S. Cordier, L. Pareschi, G. Toscani. On a kinetic model for a simple market economy. *Journal of Statistical Physics*, 120(1–2): 253–277, 2005.
45. I. D. Couzin, J. Krause, N. R. Franks, S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025): 513–516, 2005.
46. E. Cristiani, B. Piccoli, A. Tosin. Multiscale modeling of granular flows with application to crowd dynamics. *Multiscale Modeling & Simulation*, 9(1): 155–182, 2011.
47. N. Crokidakis. Role of noise and agents' convictions on opinion spreading in a three-state voter-like model. *Journal of Statistical Mechanics: Theory and Experiment*, 07: P07008, 2013.
48. N. Crokidakis, C. Anteneodo. Role of conviction in nonequilibrium models of opinion formation. *Physical Review E*: 86(6): 061127, 2012.
49. F. Cucker, S. Smale. Emergent behavior in flocks. *IEEE Transaction on Automatic Control*, 52(5): 852–862, 2007.
50. A. Das, S. Gollapudi, K. Munagala. *Modeling opinion dynamics in social networks*, Proceedings of the 7th ACM international conference on Web search and data mining, ACM New York, 403–412, 2014.
51. G. Deffuant, F. Amblard, G. Weisbuch, T. Faure. How can extremism prevail? A study based on the relative agreement interaction model. *Journal of Artificial Societies and Social Simulation*, 5(4), 2002.
52. P. Degond, M. Herty, J-G Liu, Meanfield games and model predictive control. *arXiv preprint*, 2014. [arXiv:1412.7517](https://arxiv.org/abs/1412.7517)
53. P. Degond, S. Motsch. Continuum limit of self-driven particles with orientation interaction. *Mathematical Models and Methods in Applied Sciences*, 18: 1193–1215, 2008.
54. P. Degond, J.-G. Liu, S. Motsch, V. Panferov. Hydrodynamic models of self-organized dynamics: derivation and existence theory. *Methods and Applications of Analysis*, 20(2): 89–114, 2013.
55. P. Degond, J.-G. Liu, C. Ringhofer. Large-scale dynamics of mean-field games driven by local Nash equilibria. *Journal of Nonlinear Science*, 24(1): 93–115, 2014.
56. M. Dolfin, L. Miroslav. Modeling opinion dynamics: how the network enhances consensus. *Networks & Heterogeneous Media*, 10(4): 877–896, 2015.
57. M. D'Orsogna, Y. L. Chuang, A. Bertozzi, L. Chayes. Self-propelled particles with soft-core interactions. Patterns, stability and collapse. *Physical Review Letters*, 96: 104302, 2006.
58. B. Düring, M.-T. Wolfram. Opinion dynamics: inhomogeneous Boltzmann-type equations modeling opinion leadership and political segregation. *Proceedings of the Royal Society of London A*, 471(2182): 20150345, 2015.
59. B. Düring, P. A. Markowich, J.-F. Pietschmann, M.-T. Wolfram. Boltzmann and Fokker-Planck equations modeling opinion formation in the presence of strong leaders. *Proceedings of the Royal Society of London A*, 465(2112): 3687–3708, 2009.
60. M. Fornasier, F. Solombrino. Mean-field optimal control. *ESAIM: Control, Optimisation and Calculus of Variations*, 20(4): 1123–1152, 2014.
61. M. Fornasier, J. Haskovec, G. Toscani. Fluid dynamic description of flocking via Povzner-Boltzmann equation. *Physica D: Nonlinear Phenomena*, 240(1): 21–31, 2011.
62. M. Fornasier, B. Piccoli, F. Rossi. Mean-field sparse optimal control, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 372(2028): 20130400, 21, 2014.

63. G. Furioli, A. Pulvirenti, E. Terraneo, G. Toscani. The grazing collision limit of the inelastic Kac model around a Lévy-type equilibrium. *SIAM Journal of Mathematical Analysis*, 44: 827–850, 2012.
64. S. Galam, J. D. Zucker. From individual choice to group decision-making. *Physica A: Statistical Mechanics and its Applications*, 287(3–4): 644–659, 2000.
65. S. Galam, Y. Gefen, Y. Shapir. Sociophysics: a new approach of sociological collective behavior. *Journal of Mathematical Sociology*, 9: 1–13, 1982.
66. J. Gómez-Serrano, C. Graham, J.-Y. Le Boudec. The bounded confidence model of opinion dynamics. *Mathematical Models and Methods in Applied Sciences*, 22(02): 1150007, 2012.
67. S. Y. Ha, E. Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *Kinetic and Related Models*, 1: 415–435, 2008.
68. D. Helbing, S. Lämmer, J.-P. Lebacque. Self-organized control of irregular or perturbed network traffic. *Optimal Control and dynamic games*, Springer US: 239–274, 2005.
69. M. Herty, C. Ringhofer. Feedback controls for continuous priority models in supply chain management. *Computational Methods in Applied Mathematics*, 11(2): 206–213, 2011.
70. M. Herty, C. Ringhofer. Averaged kinetic models for flows on unstructured networks. *Kinetic and Related Models*, 4: 1081–1096, 2011.
71. M. Herty, M. Zanella. Performance bounds for the mean-field limit of constrained dynamics. *Discrete and Continuous Dynamical Systems A*, 37(4): 2023–2043, 2017.
72. R. Hegselmann, U. Krause. Opinion dynamics and bounded confidence, models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.
73. M. Kristic, I. Kanellakopoulos, P. Kokotovic. *Nonlinear and Adaptive Control Design*, John Wiley and Sons Inc., New York, 1995.
74. M. Lallouache, A. Chakrabarti, A. Chakraborti, B. K. Chakrabarti. Opinion formation in the kinetic exchange models: spontaneous symmetry breaking transition. *Physical Review E*, 82: 056112, 2010.
75. P.F. Lazarsfeld, B.R. Berelson, H. Gaudet. The people's choice: how the voter makes up his mind in a presidential campaign. New York, NY: Duell, Sloan & Pierce 1944.
76. E. W. Larsen, C. D. Levermore, G. C. Pomraning, J. G. Sanderson. Discretization methods for one-dimensional Fokker-Planck operators. *Journal of Computational Physics*, 61: 359–390, 1985.
77. J.-M. Lasry, P.-L. Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1): 229–260, 2007.
78. T. Lux, M. Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719): 498–500, 1999.
79. D. Maldarella, L. Pareschi. Kinetic models for socio-economic dynamics of speculative markets. *Physica A: Statistical Mechanics and its Applications*, 391(3): 715–730, 2012.
80. D.Q. Mayne, H. Michalska. Receding horizon control of nonlinear systems. *IEEE Transactions on Automatic Control*, 35(7): 814–824, 1990.
81. D.Q. Mayne, J.B. Rawlings, C.V. Rao, P.O.M. Scokaert. Constrained model predictive control: stability and optimality. *Automatica*, 36(6): 789–814, 2000.
82. H. Michalska, D.Q. Mayne. Robust receding horizon control of constrained nonlinear systems. *IEEE Transactions on Automatic Control*, 38(11): 1623–1633, 1993.
83. M. Mohammadi, A. Borzi. Analysis of the Chang-Cooper discretization scheme for a class of Fokker-Planck equations. *Journal of Numerical Mathematics*, 23(3): 271–288, 2015.
84. S. Motsch, E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4): 577–621, 2014.
85. C. Mudde. *Populist radical right parties in Europe*. Cambridge, UK: Cambridge University Press, 2007.
86. G. Naldi, L. Pareschi, G. Toscani. *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, Birkhauser, Boston, 2010.
87. M.E.J. Newman. The structure and function on complex networks. *SIAM Review*, 45(2): 167–256, 2003.

88. L. Pareschi, G. Russo. An introduction to Monte Carlo methods for the Boltzmann equation. *ESAIM: Proceedings*, EDP Sciences. Vol. 10: 35–75, 2001.
89. L. Pareschi, G. Toscani. *Interacting Multiagent Systems. Kinetic Equations and Monte Carlo Methods*. Oxford University Press, 2013.
90. L. Pareschi, G. Toscani. Wealth distribution and collective knowledge: a Boltzmann approach. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 372(2028): 20130396, 2014.
91. L. Pareschi, P. Vellucci, M. Zanella. Kinetic models of collective decision-making in the presence of equality bias. *Physica A: Statistical Mechanics and its Application*, 467: 201–217, 2017.
92. S. Patterson, B. Bamieh. *Interaction-driven opinion dynamics in online social networks*, Proceedings of the First Workshop on Social Media Analytics, ACM New York, 98–110, 2010
93. H. Risken, *The Fokker-Planck equation*, vol. 18 of Springer Series in Synergetics, Springer-Verlag, Berlin, second ed., 1989. Methods of solution and applications.
94. P. Sen. Phase transitions in a two-parameter model of opinion dynamics with random kinetic exchanges. *Physical Review E*, 83(1): 016108, 2011.
95. E.D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*, Springer Science, Vol. 6, Second Edition, 1998.
96. S.H. Strogatz. Exploring complex networks. *Nature*, 410(6825): 268–276, 2001.
97. K. Szajd-Weron, J. Szajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(6): 1157–1165, 2000.
98. G. Toscani. Kinetic models of opinion formation. *Communications in Mathematical Sciences*, 4(3): 481–496, 2006.
99. F. Vazquez, P. L. Krapivsky, S. Redner. Constrained opinion dynamics: freezing and slow evolution. *Journal of Physics A: Mathematical and General*, 36(3): L61, 2003.
100. T. Vicsek, A. Zafeiris. Collective motion. *Physics Reports*, 517(3): 71–140, 2012.
101. C. Villani. On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations. *Archive for Rational Mechanics and Analysis*, 143(3): 273–307, 1998.
102. D.J. Watts, S.H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393: 440–442, 1998.
103. W. Weidlich. *Sociodynamics: a Systematic Approach to Mathematical Modeling in the Social Sciences*, Harwood Academic Publishers, Amsterdam, 2000.
104. G. Weisbuch, G. Deffuant, F. Amblard. Persuasion dynamics. *Physica A: Statistical Mechanics and its Applications*, 353: 555–575, 2005.
105. Y.-B. Xie, T. Zhou, B.-H. Wang. Scale-free networks without growth. *Physica A: Statistical Mechanics and its Applications*, 387: 1683–1688, 2008.

Interaction Network, State Space, and Control in Social Dynamics

Aylin Aydoğdu, Marco Caponigro, Sean McQuade, Benedetto Piccoli, Nastassia Pouradier Duteil, Francesco Rossi and Emmanuel Trélat

Abstract In the present chapter, we study the emergence of global patterns in large groups in first- and second-order multiagent systems, focusing on two ingredients that influence the dynamics: the interaction network and the state space. The state space determines the types of equilibrium that can be reached by the system. Meanwhile, convergence to specific equilibria depends on the connectivity of the interaction network and on the interaction potential. When the system does not satisfy the necessary conditions for convergence to the desired equilibrium, control can be exerted, both on finite-dimensional systems and on their mean-field limit.

A. Aydoğdu · S. McQuade · B. Piccoli · N. Pouradier Duteil (✉)

Rutgers University, Camden, NJ, USA

e-mail: nastassia.pouradierduteil@rutgers.edu

S. McQuade

e-mail: sean.mcquade@rutgers.edu

B. Piccoli

e-mail: piccoli@camden.rutgers.edu

A. Aydoğdu

e-mail: aylinvet87@gmail.com

M. Caponigro

Conservatoire National des Arts et Métiers, Equipe M2N, 292 rue Saint-Martin,
75003 Paris, France

e-mail: marco.caponigro@cnam.fr

F. Rossi

Aix Marseille Université, CNRS, ENSAM, Université de Toulon,
LSIS UMR 7296, 13397 Marseille, France
e-mail: francesco.rossi@lsis.org

E. Trélat

Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR,
7598 Paris, France
e-mail: emmanuel.trelat@upmc.fr

E. Trélat

Laboratoire Jacques-Louis Lions, Institut Universitaire de France,
75005 Paris, France

1 Introduction

A fascinating feature of large groups of autonomous agents is their ability to form organized global patterns even when individual agents interact only at a local scale. This is usually referred to as *self-organization*. We use the term *social dynamics* to indicate the study of such global behaviors, with an emphasis on understanding the mechanisms leading from local rules to global phenomena, as well as identifying the resulting global pattern formation.

Social dynamics models can be classified as first-order models and second-order models. In first-order models, we refer to the variables of interest as *opinions*, even though such models can describe a wide range of attributes such as positions, market shares, or wealth. The opinion of each agent is affected by neighboring agents' opinions in the state space. On the other hand, in second-order models, the variables of interest are the *velocities*, obtained as the time derivatives of the positions. Each agent's velocity is affected by the velocities of agents whose positions are close in the state space.

First-order models (or opinion dynamics) can give rise to patterns such as *consensus* (i.e., agreement of all states), *polarization* (i.e., disagreement between two opposite parties), or *clustering* (i.e., breakdown of the opinions into several subsets). A first formulation of opinion dynamics can be traced back to French's research on social influence [36], followed by works by Harary [42], De Groot [24], and Lehrer [60], all focusing on linear models. More recently, nonlinear models were introduced and analyzed by Krause [55, 56], Dittmer [30], Hegselmann, and Flache [44].

Second-order models are commonly applied to animal groups to study coordinated collective behavior (as done by Couzin et al. [20], Cristiani, Frasca and Piccoli [21], Giardina [37], Krause and Ruxton [54], Leonard [61], and Sumpter [82]), for example, in fish (Huth and Vissel [49], Parrish, Viscido and Grunbaum [68]) or birds (Ballerini et al. and Cucker and Smale [6, 23]). Some models have been designed to include simple interaction rules such as attraction, short-distance repulsion, and mimetic orientation or alignment. Agreement of all agents in the velocity variable is referred to as *alignment* or *flocking*.

The aim of this survey is to describe the role of two elements affecting the dynamics for both first-order and second-order models: the interaction network and the state space. We will also explore ways to control the dynamics to drive the system to a desired state.

The interaction network plays a critical role in the emergence of global patterns. Depending on the network, opinion formation models may lead to consensus among all the opinions or to the formation of separate clusters. The system's dynamics and the network's dynamics may be coupled. For instance, bounded confidence models allow agents to interact only if they are within a certain radius of each other in the state space, as proposed by Hegselmann and Krause [45]. On the other hand, it was shown that heterophilious dynamics enhances consensus by Motsch and Tadmor [65]. Another distinction can be made between metric and topological interactions. A network based on metric interactions links agents based on their distance in the state

space, whereas one based on topological interactions links an agent to another if it is among its k closest neighbors, which can lead to asymmetric relations and interesting patterns. Furthermore, a network may be constant in time or time-dependent.

The state space is another factor that greatly influences the dynamics. Most studies have considered dynamics in Euclidean spaces (most often one-dimensional for opinion models and two- or three-dimensional for animal groups). One can also study the same dynamics on general Riemannian manifolds. For instance, a nonlinear model of opinion formation on the sphere was studied by Caponigro, Lai, and Piccoli [16], with a rich structure leading to unusual equilibria. These models are based on the projection of the linear dynamics in the ambient space onto the tangent space of the manifold. Consensus dynamics on special orthogonal groups were also studied, for example, by Sarlette and Sepulchre [73, 76], motivated by applications to satellites or ground vehicles.

A large number of applications involve control of robotic networks or autonomous vehicles, as done by Bullo, Cortés, and Martínez [11]. Control is used to impose consensus or alignment when it is not reached naturally (see Caponigro, Fornasier, Piccoli, and Trélat [14, 15]), or to guide the agents in a specific direction, as done by Leonard for the migration of animal groups [61]. Ways of controlling the system include spreading leaders among the group or acting on the network. Due to the high dimensionality of social dynamics systems, control can be excessively demanding in computational resources. It is then convenient to consider the mean-field limit of the system. Numerous theories have been developed to control the resulting kinetic equation. Some approaches require taking the limit (in some sense) of the finite-dimensional controlled system. For example, Fornasier and Solombrino have introduced a concept of Γ -limit for optimal control problems [35], and Fornasier, Piccoli, and Rossi have extended the idea of control by leaders [34]. One can also control the PDE directly, as done by Piccoli, Rossi, and Trélat [71]. Other approaches involve controlling the interaction kernel (see Albi, Herty, and Pareschi [2]), or using mean-field games, a theory developed by Lasry and Lions [59] and Caines [13].

2 Overview of Social Dynamics Problems

In this section, we give general definitions of the concepts that we will use. We also provide some examples of common first-order and second-order systems and distinguish between finite-dimensional and infinite-dimensional models.

2.1 General Notations and Definitions

In the following chapter, we will differentiate between two branches of models:

- First-order models (also referred to as opinion dynamics) that can lead to *consensus*

- Second-order models (mostly related to animal group models) that can lead to *flocking* or *alignment*

We shall write first-order dynamics as follows:

$$\dot{x}_i = \frac{1}{N} \sum_{j \in \mathcal{N}_i} a_{ij} (x_j - x_i), \quad i \in \{1, \dots, N\}, \quad (1)$$

and second-order systems as follows:

$$\begin{cases} \dot{x}_i = v_i \\ \dot{v}_i = \frac{1}{N} \sum_{j \in \mathcal{N}_i} a_{ij} (v_j - v_i) \end{cases} \quad i \in \{1, \dots, N\}, \quad (2)$$

where N is the number of agents, $x_i \in \mathbb{R}^d$ is the position of agent i in the state space, $v_i \in \mathbb{R}^d$ is its velocity, \mathcal{N}_i is the set of agents interacting with agent i , and a_{ij} are interaction coefficients for each pair of agents (i, j) . They form the interaction matrix $A = (a_{ij})_{i,j \in \{1, \dots, N\}}$. Unless otherwise specified, we consider that first-order systems evolve in \mathbb{R}^{Nd} (where d is the dimension of the state space) and second-order systems are in \mathbb{R}^{2Nd} .

Remark 1 The dynamics (1) and (2) can be written in a more general form: $\dot{x}_i = \frac{1}{\deg_i} \sum_{j \in \mathcal{N}_i} a_{ij} (x_j - x_i)$ or $\dot{x}_i = v_i$; $\dot{v}_i = \frac{1}{\deg_i} \sum_{j \in \mathcal{N}_i} a_{ij} (v_j - v_i)$, where \deg_i is a scaling factor. Typical choices for scaling factors are as follows: $\deg_i = N$, $\deg_i = \text{card}(\mathcal{N}_i)$ or $\deg_i = \sum_{j \in \mathcal{N}_i} a_{ij}$, where $\text{card}(\cdot)$ denotes the cardinality of a set.

First-order models are also referred to as *consensus* models. Consensus is an equilibrium state in which all agents have the same opinion: $x_i = x_j$ for all $i, j \in \{1, \dots, N\}$. Second-order models are also called *alignment* (or *flocking*) models. Alignment or flocking is the equilibrium set in which all agents have the same velocity: $v_i = v_j$ for all $i, j \in \{1, \dots, N\}$. For this reason, the velocity v is also referred to as the *consensus variable*, to distinguish from the position x .

The system can be viewed as a network represented by a (possibly time-varying) directed weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We define the set of vertices $\mathcal{V} = (v_i)_{i \in \{1, \dots, N\}}$ corresponding to the set of agents, and the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, so that an edge exists between two vertices i and j if and only if $a_{ij} \neq 0$. The edges are weighted by the interaction coefficients a_{ij} .

Most often, the interaction coefficients are defined by an interaction potential $a(\cdot)$ such that $a_{ij} := a(\|x_i - x_j\|)$. When modeled as such, the strength of interaction is a function of the distance between agents in the position space. This generates a fundamental difference between first-order and second-order models. In first-order models, the variable of interest is the position and its tendency to agree with other agents' positions depends on the distance between the agents. In second-order models, the velocity's tendency to align with other agent's velocities depends on the difference in their positions.

If the interactions between agents are only local, that is, if an agent interacts exclusively with close neighbors, we refer to *bounded confidence* models, a term first introduced by Hegselmann and Krause [45]. Then, \mathcal{N}_i denotes the set of closest neighbors of the i -th agent. Bounded confidence models will be examined in Section 3.1.2. We look in particular at two ways to define proximity of agents. In the case of bounded confidence with metric interaction, given a radius $r > 0$,

$$\mathcal{N}_i^r(x) = \{j \in \{1, \dots, N\}, \|x_i - x_j\| \leq r\} \quad (3)$$

In the case of bounded confidence with topological interaction, we define the relative separation between two agents as $\alpha_{ij} = \text{card}\{k : \|x_i - x_k\| \leq \|x_i - x_j\|\}$. Then, the set of neighbors of agent i is defined as the set of its k closest neighbors, i.e.,

$$\mathcal{N}_i^k(x) = \{j \in \{1, \dots, N\}, \alpha_{ij} \leq k\}, \quad (4)$$

for a given $k \in \mathbb{N}$.

When all agents interact with all others, $\mathcal{N}_i = \{1, \dots, N\}$ for all $i \in \{1, \dots, N\}$. Then, the network is fully connected but its edges may have varying weights. The behavior of the system will depend on the interaction potential $a(\cdot)$, as seen in Section 3.2.

2.2 Examples of First-Order Consensus Models

We start by giving two examples of common first-order consensus models. The *Voter* model is a discrete-time model, whereas the Hegselmann–Krause model is a system of ODEs.

2.2.1 The Voter Model

One system used to explore the dynamics of cellular automata is the Sznajd model (SM) [9]. The SM is an example of discrete-time and discrete-state model. The alignment variable of each agent can take one of the two values, referred to as *spin up* or *spin down*. The dynamics of this particular model operate on a one- or two-dimensional lattice. In this system, the agents change their opinion (spin up or spin down) based on specific interaction rules: the *ferromagnetic* interaction (that is, if $x_i = x_{i+1}$, then at the next step, adjacent agents will satisfy with a given probability $x_{i-1} = x_i = x_{i+1} = x_{i+2}$) and the *antiferromagnetic* interaction (if $x_i = -x_{i+1}$, then an antisymmetric pattern forms: $-x_{i-1} = x_i = -x_{i+1} = x_{i+2}$). The model has been extended to higher dimensional opinion and complex network topologies. The motivation for this model comes from the postulate that “agreement generates agreement,” that is, if two agents reach a consensus, then all agents directly connected

with them are induced to agree. In other words, in the Sznajd model, the opinion flows out from a group of agreeing agents, a concept known as *social validation*.

In [9], the authors show that the SM is a special case of a linear Voter model. The Voter model (VM) is one of the simplest mathematical models of cooperative behavior, and its dynamics are well understood. Here, each node of a graph begins as either one of the two states: spin up or spin down. The system then follows an algorithm:

1. pick a random voter
2. the selected voter adopts the state of a randomly chosen neighbor
3. repeat steps 1 and 2 until consensus

Once the system reaches consensus, all nodes are spin up or spin down. In this system, the interaction between a randomly selected voter x_i and a randomly chosen neighbor x_j is $x_i = x_j$. In other words, the interaction is described by complete agreement with one of the neighboring agents.

2.2.2 The Hegselmann–Krause Model

The Hegselmann–Krause model (HK) is a classical example of a first-order nonlinear opinion formation model [45]. It was designed in the context of opinion dynamics and captures well-known phenomena such as the formation of consensus and emergence of clustering. Agents modify their own opinion to average neighboring opinions as follows:

$$\dot{x}_i = \frac{1}{\text{card}(\mathcal{N}_i)} \sum_{j \in \mathcal{N}_i} (x_j - x_i) \quad \text{for all } i \in \{1, \dots, N\}, \quad x_i \in \mathbb{R}^d, \quad (5)$$

where $\mathcal{N}_i = \{j : \|x_i - x_j\| \leq r\}$, $r > 0$, is the set of agents interacting with agent i . The radius r can be interpreted as the level of confidence. This model captures the fact that an individual tends to trust only opinions that do not differ from its own by more than r . Since the interaction region is bounded, the HK model is also called *bounded confidence* model. Depending on the size of the interaction regions and the density of agents in the domain, different phenomena are observed. If the interaction is strong enough (i.e., r is big enough), the agents can be brought to consensus, i.e., convergence to a single opinion. If the interaction regions are too restricted, one observes clustering around different opinions. A wide variety of models have been developed by varying the confidence region \mathcal{N}_i . Hegselmann and Krause have, for instance, looked at (one-dimensional) asymmetric confidence: $\mathcal{N}_i = \{j : -r_l \leq x_i - x_j \leq r_r\}$, $r_l > 0$, $r_r > 0$ [45]. Recently, Motsch and Tadmor have analyzed models with interaction strength increasing with the distance between agents, showing that this so-called heterophilous dynamics enhances consensus [65].

Jabin and Motsch have studied a slightly different model for opinion formation that can be written as follows:

$$\dot{x}_i = \frac{\sum_j \phi_{ij} (x_j - x_i)}{\sum_j \phi_{ij}}, \quad i \in \{1, \dots, N\}, \quad x_i \in \mathbb{R}^d, \quad (6)$$

where ϕ is the influence function and we define $\phi_{ij} := \phi(\|x_i - x_j\|^2)$ [51]. One can prove that under appropriate conditions on the influence function, the system leads to clustering. For instance, if

- $\phi \in L^\infty(\mathbb{R}^d)$ with compact support in $[0, 1]$
- for any $\varepsilon > 0$, $\phi \in W^{1,\infty}([0, 1 - \varepsilon])$ and ϕ is strictly positive on $[0, 1 - \varepsilon]$
- $|\phi'(r)|^2 \leq C\phi(r)$ for all $r \in [0, 1]$

then there exists a set of cluster centers $\{\bar{x}_i\}$ such that for all i , $x_i(t) \rightarrow_{t \rightarrow \infty} \bar{x}_i$, and for any i, j , either $\bar{x}_i = \bar{x}_j$ or $|\bar{x}_i - \bar{x}_j| \geq 1$ [51].

In one dimension, we can even characterize the rate of convergence to the clusters in the following way. Assume that $d = 1$ and $\phi \in W^{1,\infty}([0, 1))$ with $\inf_{[0,1)} \phi > 0$. Then, for each agent i , there exists \bar{x}_i depending on the initial positions of all the agents such that $|x_i(t) - \bar{x}_i| \leq Ce^{-\lambda(t-t_0)}$ for all $t \geq t_0$, where the constants C and λ are determined a priori by the total number of agents N and by the influence function ϕ , and the time t_0 depends on N , ϕ and the diameter of the initial support [51].

2.3 Examples of Finite-Dimensional and Infinite-Dimensional Second-Order Alignment Models

There exists a wide variety of second-order models that have been developed mainly to describe the behaviors of animal groups or robotic networks. Some early models like the Vicsek model [86] are defined in discrete time and require to update each agent's state at successive time intervals. Other models like the Cucker–Smale one [23] are continuous in time and require the use of ODEs. We also look at the limit of such models when the number of agents tends to infinity, which is referred to as the *mean-field limit*.

2.3.1 The Vicsek Model

A classic example of discrete-time model is the Vicsek model [86], proposed to describe interactions within animal groups such as a school of fish. It represents each agent (or fish) by its position x_k and its velocity angle θ_k , all velocities having constant norm v . The positions and angles are updated in the following way:

$$\begin{cases} x_k(t + \Delta t) = x_k(t) + v_k(t) \Delta t \\ \theta_k(t + \Delta t) = \langle \theta(t) \rangle_r + \Delta \theta_k \end{cases} \quad (7)$$

where $\Delta\theta_k$ is a noise term, and $\langle \theta(t) \rangle_r$ represents the average direction of the velocities of particles being within a circle of radius r of particle k .

2.3.2 The Finite-Dimensional Cucker–Smale Model

The prototypical second-order model for the interaction of N agents is the *Cucker–Smale model* (CS) [23]:

$$\begin{cases} \dot{x}_i(t) &= v_i(t) \\ \dot{v}_i(t) &= \frac{1}{N} \sum_{j=1}^N a(\|x_j(t) - x_i(t)\|)(v_j(t) - v_i(t)), \end{cases} \quad i = 1, \dots, N \quad (8)$$

where $x_i \in \mathbb{R}^d$, $v_i \in \mathbb{R}^d$, and $a \in C^1([0, +\infty))$ is a nonincreasing positive function called interaction potential or rate of communication. In the classical CS model, we have $a(s) = \frac{1}{(1+s^2)^\beta}$, with $\beta > 0$. Here, x_i is the main state of the agent i , and v_i is its consensus parameter. This model was initially introduced to describe the formation and evolution of languages, and was then also used for describing the flocking of a swarm of birds [23] or spacecraft formation [69].

We consider the finite-dimensional CS model (8) and provide some results published by Caponigro, Fornasier, Piccoli, and Trélat [14]. We define the space barycenter $\bar{x}(t)$ and the mean velocity $\bar{v}(t)$ by

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

Then, $\dot{\bar{x}}(t) = \bar{v}$ and the mean velocity is constant: $\bar{v}(t) = \bar{v}(0)$ for all t . Define the spatial variance by

$$X(t) = \frac{1}{2N^2} \sum_{i,j=1}^N \|x_i(t) - x_j(t)\|^2.$$

The velocity variance is

$$V(t) = \frac{1}{2N^2} \sum_{i,j=1}^N \|v_i(t) - v_j(t)\|^2 = \frac{1}{N} \sum_{i=1}^N \|v_i(t) - \bar{v}\|^2, \quad (9)$$

and we have

$$\dot{V}(t) = -\frac{1}{N} \sum_{i,j=1}^N a(\|x_j(t) - x_i(t)\|) \|v_i(t) - v_j(t)\|^2 \leq 0.$$

Definition 1 A solution $(x(t), v(t))$ converges to *alignment* (or *flocking*) if

- (i) there exists $X_M > 0$ such that $X(t) \leq X_M$ for every $t > 0$,
- (ii) $v_i(t) \xrightarrow[t \rightarrow +\infty]{} \bar{v}$, for every $i = 1, \dots, N$, or equivalently, $V(t) \xrightarrow[t \rightarrow +\infty]{} 0$.

Note that, since a is nonincreasing, we have $\dot{V}(t) \leq -2a(\sqrt{2N}X(t)) V(t)$, and hence, if $X(t)$ remains bounded, then $\dot{V} \leq -\alpha V$, which implies flocking. But the difficulty is that the group of agents does not necessarily remain confined, and for this reason, convergence to alignment is not guaranteed. More precisely, Ha, Ha and Kim provided the following result [39].

Proposition 1 [39] Let $(x_0, v_0) \in (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N$ be such that

$$\sqrt{V(0)} \leq \int_{\sqrt{X(0)}}^{+\infty} a(\sqrt{2Nr}) dr.$$

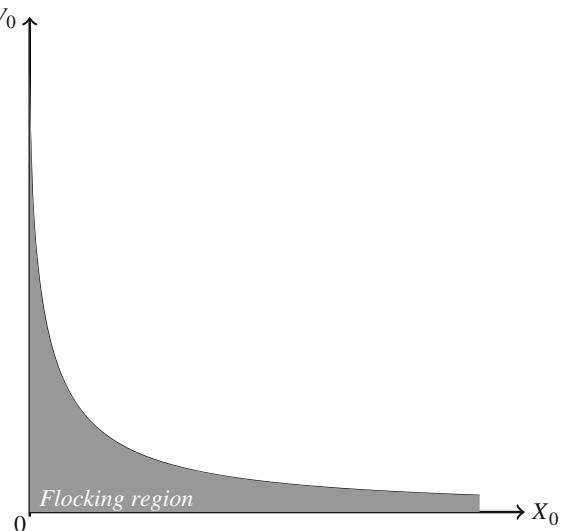
Then the solution with initial data (x_0, v_0) tends to alignment.

Figure 1 illustrates the *self-organization of the group* in what we can call the *flocking region* (region of natural asymptotic stability to flocking).

2.3.3 The Infinite-Dimensional Kinetic Cucker–Smale Model

In numerous applications such as risk-taking in economics, pricing models, and opinion formation, the system is made of a very large number of agents. Studying

Fig. 1 Flocking region
 $\sqrt{V_0} \leq \frac{1}{\sqrt{2N}}(\frac{\pi}{2} - \arctan(\sqrt{X_0}))$ corresponding to the CS model (8) with parameter $\beta = 1$, see Proposition 1



and simulating social dynamics systems becomes a particularly challenging problem when the dimension of the system increases. This is referred to as *the curse of dimensionality*, a term coined by Bellman in the context of dynamic optimization of high-dimensional systems. One way around this problem is to move away from the microscopic viewpoint where each agent is considered individually, and consider instead the mean-field limit, which provides a kinetic description of the system. This approach consists of approximating the influence of all agents on any given individual by one averaged effect. Derivation of kinetic models has been intensively studied, for example, by Cañizo, Carrillo, and Rosado [12], by Ha and Tadmor for the CS model [41], or by Degond and Motsch for the Vicsek model [28, 29].

When the number of agents is large, one often refers to the agents as *particles*. Let $\mu(t, x, v)$ denote the distribution function of particles positioned at $x \in \mathbb{R}^d$ at time $t > 0$ with velocity $v \in \mathbb{R}^d$. By taking the mean-field limit in system (8), we obtain the *kinetic Cucker–Smale* model:

$$\partial_t \mu + \langle v, \nabla_x \mu \rangle + \operatorname{div}_v (\xi[\mu] \mu) = 0 \quad (10)$$

where $\mu(t)$ is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ (if $\mu(t, x, v) = f(t, x, v) dx dv$, then f is the density of the particles), and $\xi[\mu]$ is the *interaction kernel*, given by

$$\xi[\mu](x, v) = \int_{\mathbb{R}^d \times \mathbb{R}^d} a(\|x - y\|)(w - v) d\mu(y, w).$$

The link with the finite-dimensional system is given by the empirical measure

$$\mu(t) = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i(t), v_i(t))}.$$

Indeed, plugging this measure in (10), we find that $(x_i(t), v_i(t))$ satisfy exactly (8). The kinetic equation (10) can be written as

$$\partial_t \mu + \operatorname{div}_{(x,v)} (V[\mu] \mu) = 0,$$

with the velocity field

$$V[\mu] = \begin{pmatrix} v \\ \xi[\mu] \end{pmatrix}.$$

The so-called *particle flow* $\Phi(t)$ generated by $V[\mu(t)]$ yields the *characteristics*

$$\dot{x}(t) = v(t), \quad \dot{v}(t) = \xi[\mu(t)](x(t), v(t)).$$

The motion of any such particle follows exactly the finite-dimensional CS system. This justifies the wording *particle*. Moreover, the solution to the kinetic equation (10) is formally as follows:

$$\mu(t) = \Phi(t)\#\mu^0,$$

that is, the pushforward under the flow $\Phi(t)$ of the initial measure.

Similarly to the finite-dimensional case, we present some of the properties of the infinite-dimensional model. In the infinite-dimensional setting, we define the space barycenter and mean velocity by

$$\bar{x}(t) = \int_{\mathbb{R}^d \times \mathbb{R}^d} x \, d\mu(t)(x, v), \quad \bar{v} = \int_{\mathbb{R}^d \times \mathbb{R}^d} v \, d\mu(t)(x, v).$$

Then, $\dot{\bar{x}}(t) = \bar{v}$ and $\dot{\bar{v}} = 0$, as in finite dimension. Defining (as before) the spatial and velocity variances by

$$X(t) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - \bar{x}(t)\|^2 \, d\mu(t)(x, v), \quad V(t) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|v - \bar{v}\|^2 \, d\mu(t)(x, v),$$

we have

$$\dot{V}(t) = - \iint a(\|x - y\|) \|v - w\|^2 \, d\mu(t)(x, v) \, d\mu(t)(y, w) \leq 0.$$

We expect that $V(t) \xrightarrow[t \rightarrow +\infty]{} 0$, but as in finite dimension, this is not guaranteed, unless the population of agents remains confined. This justifies the following definition. The notation supp stands for the support of a measure.

Definition 2 A solution $\mu \in C^0(\mathbb{R}, \mathcal{P}_c(\mathbb{R}^d \times \mathbb{R}^d))$ converges to *alignment* (or *flocking*) if:

- (i) there exists $X_M > 0$ such that $\text{supp}(\mu(t)) \subseteq B(\bar{x}(t), X_M) \times \mathbb{R}^d$ for every $t > 0$,
- (ii) $V(t) \xrightarrow[t \rightarrow +\infty]{} 0$.

Piccoli, Rossi, and Trélat [71] provided the infinite-dimensional counterpart of Proposition 1. As in finite dimension, it defines a consensus region, that is, a set of initial conditions for which the group of agents will naturally converge to alignment:

Proposition 2 [71] Let $\mu^0 \in \mathcal{P}_c(\mathbb{R}^d \times \mathbb{R}^d)$. Define the space and velocity barycenters $\bar{x}^0 = \int x \, d\mu^0$, $\bar{v} = \int v \, d\mu^0$ and the space and velocity support radii:

$$X^0 = \inf \left\{ X \geq 0 \mid \text{supp}(\mu^0) \subset B(\bar{x}^0, X) \times \mathbb{R}^d \right\},$$

$$V^0 = \inf \left\{ V \geq 0 \mid \text{supp}(\mu^0) \subset \mathbb{R}^d \times B(\bar{v}, V) \right\}.$$

If

$$V^0 < \int_{X^0}^{+\infty} a(2x) \, dx,$$

then the solution $\mu(t)$ with initial datum $\mu(0) = \mu^0$ converges to consensus.

3 Role of the Interaction Network

In finite-dimensional systems, the set of interacting agents can be interpreted as the vertices \mathcal{V} of a graph \mathcal{G} , and their interactions can be represented as weighted edges \mathcal{E} , as seen in Section 2.1. In all the models reviewed in Section 2, the dynamics depend on the interaction network via the interaction coefficients a_{ij} (see systems (1) and (2)). In turn, the interaction network may depend on the dynamics, for instance, when the interaction coefficients depend on the state variables: $a_{ij} = a(\|x_i - x_j\|)$. Then, the graph \mathcal{G} varies with time. In this section, we explore the influence of the network on the dynamics and vice versa.

We will look at models in which (at least initially) the set of neighbors for each agent is smaller than the set of all agents: $\text{card}(\mathcal{N}_i) < N$, so $\mathcal{E} \subsetneq \mathcal{V} \times \mathcal{V}$, such as bounded confidence models, as defined in Section 2.2.

On the other hand, some models use the complete set of agents as the interaction network, so that each agent interacts with all the others. The interaction network is then interpreted as a weighted graph, where each edge's weight is given by the interaction coefficient a_{ij} . This is also a useful representation for mean-field limits. Indeed, when the number of agents tends to infinity, the concept of graph and neighbors is lost. Instead, the interaction potential, which can be based on the relative distance between agents, can be easily transported to the mean-field setting.

3.1 Interaction Network in Bounded Confidence Models

In this section, we study the influence of the interaction network in bounded confidence models. We review known properties of such models, propose open problems concerning the equilibrium sets, and provide numerical simulations illustrating the known and conjectured properties.

3.1.1 Properties of Bounded Confidence Models

The rationale for bounded confidence models is that it is unlikely for one agent to be influenced by another one whose opinion is too far from its own. This kind of interaction gives rise to clusters of opinions (see for instance [10]). We also mention the bounded confidence model by Deffuant, see [25], in which the opinions belong to real intervals too but the pairs of interacting agents are chosen randomly.

As mentioned in Section 2.1, two main types of interaction networks have been proposed in the literature. In metric interaction networks, agents interact depending on their distance in the state space [43]: Given a confidence radius $r > 0$, we can define the interaction neighborhood \mathcal{N}_i^r (3), see Figure 2a. In topological interaction networks, agents interact depending on their relative separation. Given $k \in \mathbb{N}$, we can define the interaction neighborhood \mathcal{N}_i^k (4), see Fig 2b.

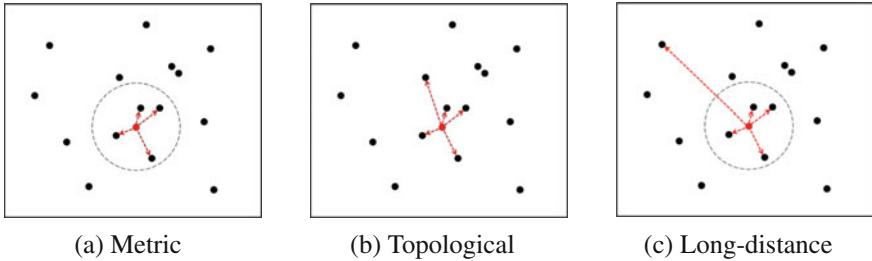


Fig. 2 Representation of interacting neighbors for one agent according to the different interaction networks. In (c), the long-distance connection is added to metric local interactions

Both topological and metric interactions are local interactions. Adding long-distance connections to local ones greatly reduces the network's diameter and facilitates the spread of information [56]. This is justified by the ubiquitous idea that social networks are of small diameter, a property also known as the six degrees of separation or *small-world effect* [88]. In particular, Kleinberg (see [53, 56]) showed that single long-distance random connections in locally organized networks lead to efficient routing procedures for spreading information. The small-world phenomenon is characterized by short paths (relative to the size of the network) connecting any two nodes in the network, as illustrated in Figure 2c. The model as presented in [53] describes nodes on a square lattice which interact with the four adjacent nodes in the lattice, as well as one long-range interaction that randomly forms an edge between a node and another non-neighboring node with a probability proportional to ρ^{-a} , where ρ is the Manhattan distance between the two nodes.

Equilibrium sets. To understand the mechanisms behind cluster formation, we studied equilibria for the HK dynamics, both with metric and topological interactions. Let us start with metric interaction, with 2 or 3 agents in \mathbb{R} :

- For $N = 2$, the equilibrium set E consists of 3 subsets: the line $x_1 = x_2$; the half-plane $x_1 - x_2 > r$; and the half-plane $x_2 - x_1 > r$ (see Figure 3a).
- In the case $N = 3$, 13 equilibrium subsets can be enumerated: the line $x_1 = x_2 = x_3$; the 3 half-planes $\{x_i = x_j, x_k > x_i + r\}$; the 3 half-planes $\{x_i = x_j, x_k < x_i - r\}$; and the 6 3D manifolds $\{x_i + r < x_j < x_k - r\}$ (with i, j, k pairwise distinct in $\{1, 2, 3\}$).

Notice that in both cases, the equilibrium set is composed of pairwise disjoint manifolds with no common boundaries. We recall the following:

Definition 3 A set $E \subset \mathbb{R}^n$ is called *stratified* in the sense of Whitney [89] if there exists a countable (locally finite) collection of pairwise disjoint manifolds $M_i, i \in \mathbb{N}$, such that the following holds:

1. M_i is an embedded manifold of dimension d_i .
2. If $M_i \cap \partial M_j \neq \emptyset$, then $M_i \subset \partial M_j$ and $d_i < d_j$.

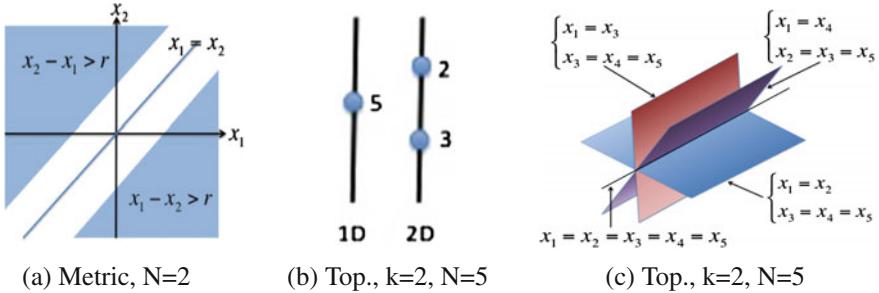


Fig. 3 Equilibria for the HK system with metric and topological interactions for $d = 1$. Figure (a) shows the equilibrium set in the metric case ($N = 2$), with separate strata. Figure (b) shows the possible configurations for the agents' positions at equilibrium for the topological interaction ($k = 2, N = 5$), indicating the number of agents in each cluster and the dimension of the manifold. Figure (c) shows some of the non-separate strata of this equilibrium set

Moreover, we say that E has separate strata if for every $i \neq j$ we have $M_i \cap \partial M_j = \emptyset$.

We propose a general property for the equilibrium set:

Conjecture 1 For the HK dynamics (5) with metric interaction, for all $d \in \mathbb{N}$ and $N \in \mathbb{N}$, the set of equilibria is a stratified manifold with separate strata.

In the topological case, the number and nature of equilibrium sets depend on k . If $k = 1$ (i.e., there is no interaction between agents), the equilibrium set is \mathbb{R}^N itself. If $k = 2$ (each agent interacts with one other), we have to distinguish cases:

- for $N = 2$ or $N = 3$, the equilibrium sets are, respectively, the lines $x_1 = x_2$ and $x_1 = x_2 = x_3$.
- for $N \geq 4$, the equilibrium sets are more complex as they are composed of several manifolds. For instance, in the case $N = 5$, the equilibrium set consists of the line $x_1 = x_2 = x_3 = x_4 = x_5$ and the $\binom{5}{2} = 10$ half-planes $\{x_i = x_j; x_k = x_l = x_m\}$ with i, j, k, l, m pairwise distinct in $\{1, \dots, 5\}$ (Figure 3b, 3c). Notice that the line is in the boundary of all half-planes.

Hence, we propose the following:

Conjecture 2 For the HK dynamics (5) with topological interaction, for any $d \geq 2$ and $N \geq 4$, the set of equilibria is a stratified manifold with non-separate strata.

3.1.2 Numerical Results

To compare the different interaction networks, we ran simulations for the well-established one-dimensional HK model, see Figure 4. Recent results [6] proposed the idea that topological interaction (with the 5-7 closest neighbors) is an effective way for birds to ensure group cohesion and to escape predators. Figure 4a, 4b show

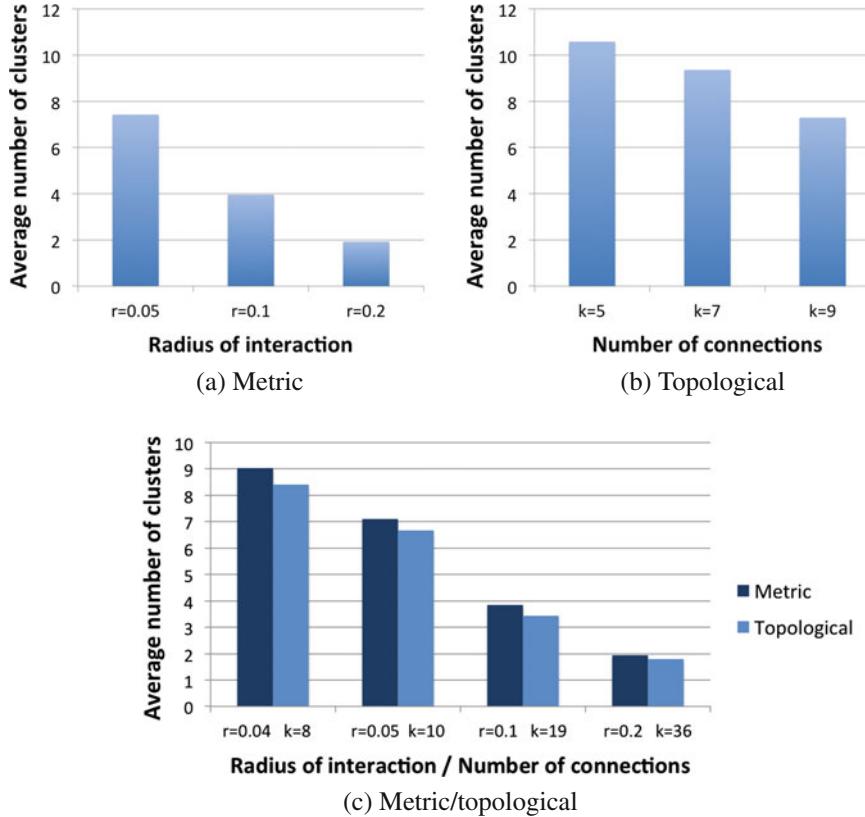


Fig. 4 Average number of clusters of the asymptotic solution: (a) for different radii r in the metric configuration, and (b) for different numbers of connections k in the topological configuration. Each average was obtained over 100 simulations, in which 100 agents are initially distributed uniformly in the interval $[0, 1]$. Figure (c) provides a comparison of the two networks, plotting side-by-side metric and topological configurations with the same initial average number of connections per agent

the average number of clusters of the asymptotic solution of the HK model (5), respectively, with metric interaction (3) and with topological interaction (4), for a group of 100 agents. Notice that consensus is not reached for small radius of interaction ($r \leq 0.2$) or a small number of neighbors ($k < 10$), but instead the group tends to cluster in several subgroups. As expected, the number of clusters decreases as the network connectivity increases. Figure 4c shows that with the same initial number of connections, both interaction networks perform similarly.

In order to illustrate the differently stable equilibrium conformations, we ran simulations with the one-dimensional HK system, plotting the distribution of the asymptotic clusters' sizes (see Figure 5). We observed that some conformations are statistically more frequent than others. For instance, in 1000 simulations of the HK dynamics of a group of 100 agents with metric interaction and an interaction

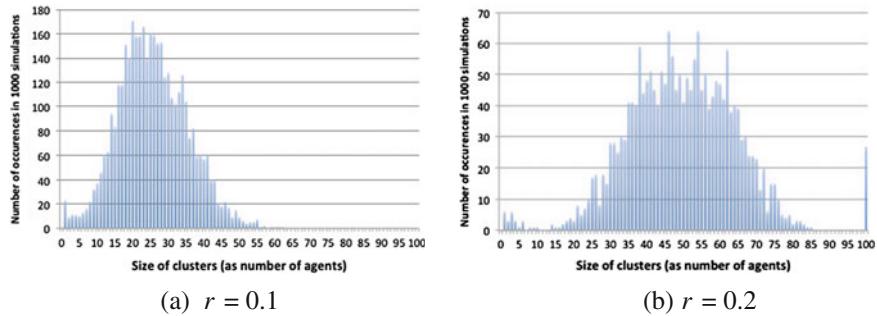


Fig. 5 Distribution of the asymptotic clusters' sizes in 1000 simulations of the one-dimensional HK model with 100 agents and metric interaction. Initially, the agents are distributed uniformly in the interval $[0, 1]$. Figure (a) was obtained with an interaction radius $r = 0.1$ and Figure (b) with $r = 0.2$. In the case $r = 0.2$, consensus was reached in 28 simulations. Furthermore, the shape of the distribution suggests that some cluster sizes are more frequent than others

radius $r = 0.2$, clusters of 38, 46, 54, and 62 agents are the most frequently obtained (Figure 5b). Notice that if $r = 0.2$ and the agents are distributed in the interval $[0, 1]$, there can be at most 4 clusters. We show that in the conditions of the simulations of Figure 5b, in most cases, the agents are asymptotically distributed in 2 clusters. Figure 6 shows the size distribution of the two biggest clusters (C_1, C_2) over 2000 simulations. The peaks are mostly distributed along the line $C_1 + C_2 = 100$, which means that in most simulations an equilibrium of 2 clusters is reached. Observe that

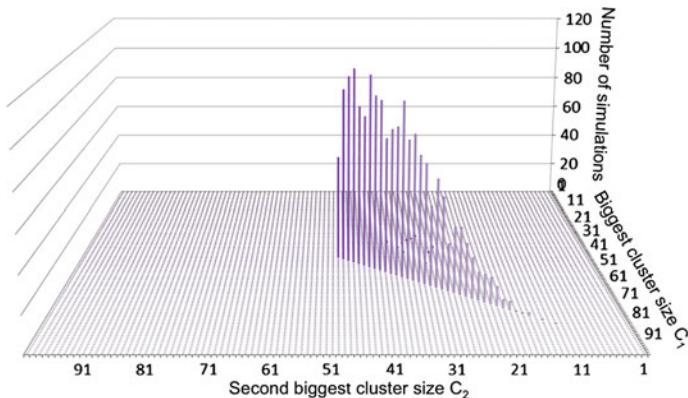


Fig. 6 Distribution of the two biggest asymptotic clusters' sizes in 2000 simulations of the one-dimensional HK model with 100 agents and metric interaction ($r = 0.2$). Initially, the agents are distributed uniformly in the interval $[0, 1]$. The conformation $(C_1, C_2) = (50, 50)$ is obtained in 60 cases, whereas the conformation $(C_1, C_2) = (51, 49)$ is obtained in 100 cases. The most likely conformation is $(C_1, C_2) = (53, 47)$, obtained in 113 cases. There is a low likelihood of having $C_1 - C_2 > 20$

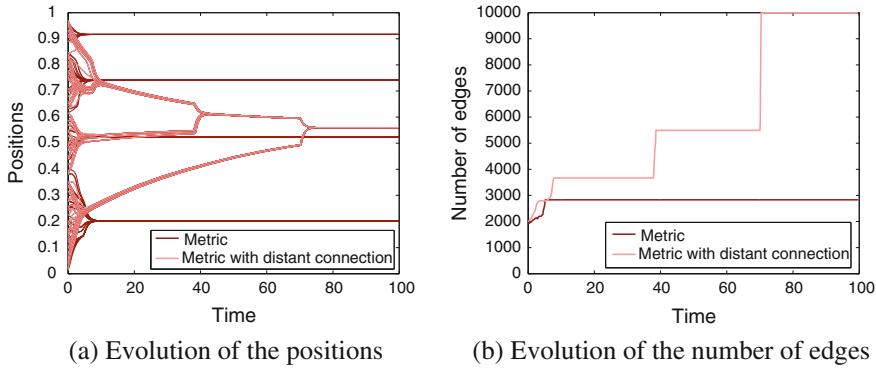


Fig. 7 Effect of distant connections in convergence to consensus in the HK model with $r = 0.1$. Figure (a) shows the evolution of positions in the metric case with only local interactions or with one added distant connection chosen uniformly (i.e., $a = 0$), resulting, respectively, in clustering or consensus. Figure (b) shows the evolution of the number of edges

it is less likely to reach an exactly equal distribution of agents between those two clusters than it is to have a slightly unbalanced distribution. The probability of having a very unbalanced distribution decreases with the imbalance.

Long-range connection We verified the effectiveness of long-distance connections in enhancing consensus for social dynamics. For each agent, a distant connection selected uniformly among the other agents was added to each agent's local interactions (see Figure 2c). Added to metric interactions, the distant connection almost always leads to consensus. Figure 7 shows the improved convergence to consensus when adding an additional distant connection in the HK model. Figure 7a shows the evolution of positions with and without an added distant connection. Figure 7b shows the evolution of the total number of edges of the network. When distant connections are added, the system asymptotically reaches consensus, and the graph becomes fully connected, i.e., $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ so that $\text{card}(\mathcal{E}) = N^2$.

We then studied the effect of the probability with which the distant connection is chosen among all the graph edges. More specifically, we penalize the increase in distance between agents by choosing the distant connection with a probability proportional to ρ^{-a} , where $a \in (0, 1)$ and ρ is the distance between agents. With local metric interaction, adding such a distant connection almost always leads to consensus. With topological interaction, consensus is not always reached but the number of final clusters is significantly reduced. The more biased the choice of distant connection is toward distant neighbors (i.e., the smaller the parameter a), the faster consensus is achieved in the metric case (Figure 8a) or the fewer clusters are obtained in the topological case (Figure 8b).

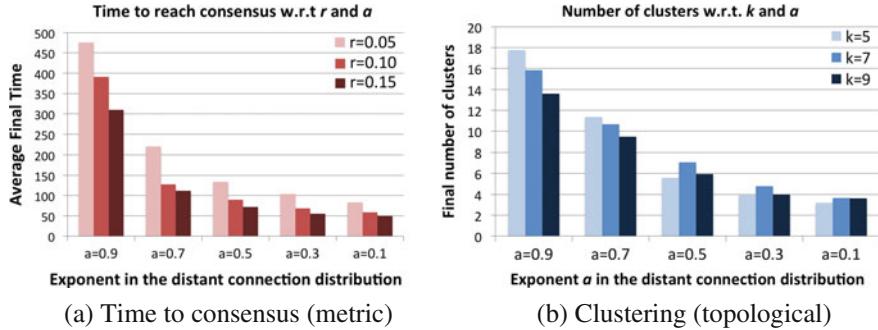


Fig. 8 Effect of distant connections in convergence to consensus in the HK model. Figure (a) shows the decrease of the time necessary to reach consensus by adding a distant connection (metric case). Since consensus is reached only asymptotically, time to consensus was defined as the time necessary for all agents to be within a sphere of given radius ε . Figure (b) shows the decrease of the final number of clusters by adding a distant connection (topological case)

3.2 The Interaction Potential

In equations (1) and (2), the interaction coefficient a_{ij} can be defined as a function of the distance between the agents i and j : $a_{ij} = a(\|x_i - x_j\|)$. The interaction potential a can be chosen to be *homophilous* or *heterophilous*, a terminology used by Motsch and Tadmor [65]. Both of these interaction potentials are functions of the distance between two agents.

- $a(\cdot)$ is a *homophilous* interaction potential if it is a decreasing function of the distance between agents.
- $a(\cdot)$ is a *heterophilous* interaction potential if it is an increasing function of the distance between agents.

A homophilous interaction potential is appropriate when aiming to model behavior that has a strong interaction between two agents that are close together, or two opinions that are similar. If a bird uses sight to maintain proximity to neighboring birds (a flock), then cohesiveness between birds will depend on distance and visual acuity. A study of nonincreasing interaction functions in the CS model is given in [23, 41]. Homophilous interaction potential is intuitive in the sense that, for agents, organisms, or opinions to influence each other, they must not be too distant.

A heterophilous interaction potential agrees with the phrase, “opposites attract.” When two agents are very different from each other, they tend to have strong influence on the other. Motsch and Tadmor show a counterintuitive result [65]: Heterophilous interaction potential increases clustering behavior. Particularly, for the long-term behavior of the system, the number of clusters of agents will decrease as the heterophilous interactions strengthen. Sufficiently strong heterophilous interactions will drive the number of clusters to one, which is a consensus.

A more detailed model implements multiple interactions, such as short-range repulsion and long-range attraction. This describes behavior where agents avoid collision but otherwise converge. In this kind of model, there is cohesion among agents, but as soon as they come too close to each other, they will move apart. This behavior is present in animal groups [54, 82] and schools of fish [49, 68], and is used to model human crowds of pedestrians [22].

Anisotropic interactions are those that depend on an orientation of an agent relative to other agents. For example, an animal may mostly receive information from its field of vision. In this case, the visual space of an agent must be considered. An animal may easily recognize animals in front, as opposed to behind. A study of how these anisotropic interactions affect the structure of animal groups can be found in [21].

4 Role of the State Space

In standard models, the agents evolve in Euclidean spaces \mathbb{R}^{Nd} or \mathbb{R}^{2Nd} . For modeling purposes, one might need to consider more complex state spaces like compact manifolds, for instance \mathbb{S}^1 or \mathbb{T} . The dynamics then give rise to new kinds of equilibria that differ from the usual consensus or alignment. We will present such models for both first-order and second-order dynamics.

4.1 First-Order Dynamics

To describe the slow and continuous evolution of opinions, several models adapted consensus algorithm on Euclidean spaces, such as the HK model described previously (Section 2.2). The dynamics of consensus models on Euclidean spaces are, at least locally, linear. This may be a limitation in representing the complex behavior of opinions. Indeed, the only equilibria of the system are clusters of consensus (see [10]). This is one of the main issues determining a lack of connections with real-life examples as pointed out by Sobkowicz [79].

Recently, there has been a growing interest in designing consensus algorithm on nonlinear manifolds. The motivation comes from engineering applications, indeed oscillators evolve on the circle S^1 , satellite altitudes evolve on the special orthogonal group $SO(3)$, and ground vehicles on the euclidean groups $SE(2)$ or $SE(3)$. The first model in this direction is the Kuramoto model [57] on the sphere \mathbb{S}^1 which attracted a wide interest of researchers over the last 30 years, motivated by its connection with the problem of synchronizing a large population of harmonic oscillators—see the survey by Strogatz [81]. Other possible applications were studied by Hopfield [47] and Vicsek et al. [86]. Lately, convergence analysis for adapted versions of the Kuramoto model on the circle has been thoroughly studied in a series of papers by Dörfler, Chertkov and Bullo [31] Scardovi, Sarlette and Sepulchre [75], Sepulchre, Paley, and Leonard [77, 78].

A first effort in studying consensus dynamics on more general manifolds has been made in [74] by Sarlette and Sepulchre who looked at, among others, the special orthogonal group $SO(n)$, the Grassmann manifold, and \mathbb{S}^1 (see also [73] and [76] for a survey on this topic). Consensus problems on general manifolds present an inherent difficulty: In order to move toward a given point (for instance, the weighted average of its neighbors' positions), an agent must follow the geodesics of the manifold, which are well defined only locally. Not only can geodesics not be unique on a global scale, but also their computation can be extremely challenging. One way around this difficulty is to consider the embedding of the manifold M into a Euclidean space E (for instance $E = \mathbb{R}^d$). Using the embedding, these models are mainly based on the projection of linear consensus dynamics on the tangent space to the manifold M . Namely, given N agents, their opinions $x_i \in M$ evolve according to the following:

$$\dot{x}_i = \Pi_{x_i} \left(\sum_{j=1}^N a_{ij} (\hat{x}_j - \hat{x}_i) \right), \quad \text{for } i = 1, \dots, N, \quad (11)$$

where \hat{x}_i denotes the embedding of x_i in E , and $\Pi_x(y)$ is the projection of y onto the tangent space to M at x . The dynamics for these systems inherit locally the structure of the linear case, and convergence results rely mainly on consensus algorithms for linear systems (as, for example, the one by Tsitsiklis [85], Jadbabaie, Lin, and Morse [52], Moreau [63, 64], Blondel, Hendrickx, Olshevsky, and Tsitsiklis [10], and Olfati-Saber and Murray [66]) but this is no longer possible for global convergence analysis since the considered manifolds are in general not globally convex. As in linear algorithms, consensus is an equilibrium of the system. In Euclidean spaces, if the interaction graph associated with the interaction coefficients a_{ij} , $i, j = 1, \dots, N$ is strongly connected, then the system always tends to consensus. In nonlinear manifolds, consensus configurations become graph-dependent.

Systems on compact manifolds show more diverse kinds of equilibrium configuration, for instance the anti-consensus, in which each state is furthest from the mean of its neighbors, so that as a result, the opinion spreads over the entire manifold. This phenomenon is sometimes called *balancing*, in opposition to the term *synchronization* used to describe consensus on the circle [75].

Recently, Caponigro, Lai, and Piccoli proposed a nonlinear opinion formation model on the d -dimensional sphere \mathbb{S}^d [16]. The rationale for the sphere \mathbb{S}^d is that, as mentioned, opinions are subjected to a quantization phenomenon when measured. We can imagine that, at the instant of measurement (elections, polls, interviews, etc.), opinions take only two values (yes/no, left/right, Democratic/Republican, liberal/conservative, for/against, etc.), so that every component of the vector $x_i = (x_i^{(1)}, \dots, x_i^{(d+1)})$ takes a positive or a negative value. In particular, x_i belongs to $\sqrt{d+1} \mathbb{S}^d$. The manifold \mathbb{S}^d is a mathematical abstraction to describe the dynamical evolution of the opinions on a continuous (i.e., non-discrete) set. Moreover, opinions on different topics are usually interconnected: economic policy attitudes and candidate choice in political elections; opinion formation and economical condition; opinion on research funding and religious or ideological beliefs.

System (11) on the sphere shows new kind configurations with respect to the one observed on Lie Groups. Beside consensus also antipodal and polygonal equilibria appear in the model (that can be seen as balanced configuration). Furthermore, a configuration typical of this system, called *dancing equilibrium*, is shown. In this configuration, the mutual distances between opinions are in equilibrium but the system may evolve.

4.2 Second-Order Dynamics

Standard second-order social dynamics systems such as the CS dynamics (8) evolve in the Euclidean space \mathbb{R}^{2Nd} , where N is the number of agents and $2d$ the dimension of the state space for the position and velocity (typically $d = 2$ or $d = 3$). However, similar to opinion dynamics models, some applications require more complicated state spaces.

One of the main difficulties in modeling opinion formations is the lack of reliable methods to measure opinions. A classical problem in sociology is to design interviews not affecting opinions, i.e., questions not influencing answers. Purely open questions do not exist and, moreover, it is very hard to collect data from open answers. On the other hand, closed questions induce quantization on the answers: Opinions collapse on discrete sets representing the possible answers to a closed question. This is the rationale to design models in which the initial and final opinions, in an opinion formation process, take value in a discrete set as in well-known Sznajd model and Voter Model, for instance. As seen in Section 2.3, the Sznajd model belongs to the class of binary-state opinion dynamics model. It is based on the Ising model for ferromagnetism in statistical mechanics [83]. In this model, opinions are discrete variables x_i taking value in the space $\{-1, 1\}$. It has been established that there are two possible equilibria for this model: *ferromagnetism*, in which all agents have the same spin, and *antiferromagnetism*, in which agents have alternate spins.

Another classic example is the Vicsek model [86] in which every particle's velocity is assumed to have constant norm, so that each particle is represented by its two-dimensional position and the angle of its velocity. This model allows to study clustering and orientational order, two patterns commonly observed in various biological systems such as animal groups or bacteria. As a variation upon the Vicsek model, Motsch and Degond designed the persistent turning walker model in order to study fish motion, where the velocity is also assumed to have constant norm c [29]. The variables are the two-dimensional position of the fish's centroid $x \in \mathbb{R}^2$, the velocity angle $\theta \in \mathbb{R}/2\pi\mathbb{Z}$ and the curvature of the trajectory $\kappa \in \mathbb{R}$. The trajectories are described by the stochastic differential equations:

$$\begin{cases} \dot{x} = c\tau(\theta) \\ \dot{\theta} = c\kappa \\ d\kappa = -a\kappa dt + b dB_t \end{cases}$$

where $\tau(\theta) = (\cos \theta, \sin \theta)$ is the direction of the velocity vector, dB_t is the standard Brownian motion, a is a relaxation frequency, and b quantifies the intensity of the random curvature jumps. The dynamics of the curvature of the trajectory reflect the antagonistic effects of its tendency to relax to a straight line and of the random jumps observed in fish behavior.

5 Control

Many works have explored ways of controlling social dynamics systems. Control can be of great use, for instance in applications to robotics, for rendezvous problems. One can aim to control the system to reach consensus in the state space or alignment in the velocity space [14, 15], reach a predetermined desired position or velocity [70], and keep the agents as far from each other as possible, to avoid Black Swan type phenomena where consensus can lead to market collapse (in applications to economics).

Control is a particularly challenging problem due to the high dimensionality of the systems. One can either control the high-dimensional discrete system [70], resort to mean-field control [34, 46], and act on the network to exploit its intrinsic properties (such as symmetry) [72]. Control often leads to separating the group into a set of controlled leaders and a set of uncontrolled followers.

5.1 Finite Dimension

We start by presenting various control techniques related to finite-dimensional models.

5.1.1 Consensus Protocols

Since the first 2000s, consensus in multiagents systems has been seen also as a distributed control problem (see for instance Jadbabaie, Lin and Morse [52], Olfati-Saber, Fax and Murray [66] and also Tsitsiklis [85]). The problem in this framework is to find a feedback control, or *consensus protocol*, $u(x) = (u_1(x), \dots, u_N(x))$ assigning dynamics to the i -th agent

$$\dot{x}_i = u_i(x)$$

for $i = 1, \dots, N$ such that every solution tends to a consensus configuration $x_1 = \dots = x_N$. The first results in this direction deal with linear consensus algorithm of the form

$$u_i(x) = \sum_{j=1}^N a_{ij}(x_j - x_i),$$

and show sufficient conditions guaranteeing asymptotic consensus under minimal connectivity assumptions on the communication graph associated with the interaction coefficients a_{ij} , $i, j = 1, \dots, N$ (see for instance Moreau [63, 64]).

5.1.2 Non-consensus

A great interest has been given to the modeling of emergent behavior in animal groups and social dynamics (See Section 2). However, self-organization is not always sufficient to ensure consensus of positions or alignment of velocities. The following example shows initial conditions for which the Cucker–Smale system does not tend to alignment.

Remark 2 Consider the CS system (8) in the case of two agents moving in \mathbb{R} with position and velocity at time t , $(x_1(t), v_1(t))$ and $(x_2(t), v_2(t))$. Assume that $a(x) = 2/(1 + x^2)$. Let $x(t) = x_1(t) - x_2(t)$ be the relative main state and $v(t) = v_1(t) - v_2(t)$ be the relative flocking parameter. Then, (8) reads

$$\begin{cases} \dot{x} = v \\ \dot{v} = -\frac{v}{1+x^2} \end{cases}$$

with initial conditions $x(0) = x_0$ and $v(0) = v_0 > 0$. The solution of this system can be found by direct integration, as from $\dot{v} = -\dot{x}/(1+x^2)$, we have

$$v(t) - v_0 = -\arctan x(t) + \arctan x_0.$$

Whenever the initial conditions satisfy $|\arctan x_0 + v_0| > \pi/2$, which implies $|\arctan x_0 + v_0| \geq \pi/2 + \varepsilon$ for some $\varepsilon > 0$, the flocking parameter $v(t)$ remains bounded away from 0 for every time, since

$$|v(t)| = |- \arctan x(t) + \arctan x_0 + v_0| \geq |-\arctan x(t) + \pi/2 + \varepsilon| > \varepsilon,$$

for every $t > 0$. In other words, the system does not tend to flocking.

5.1.3 External Control

When flocking is not achieved by self-organization, it is natural to wonder whether it is possible to control the group to flocking by means of an external action. We are therefore concerned with *organization via intervention*. Since flocking is a steady configuration of the system, enforcing self-organization can be seen as an

asymptotic stabilization problem, which is classical in control theory and usually relies on Lyapunov design (see for instance Isidori [50] or Sontag [80]). Using these classical techniques, it is easy to design a stabilizing feedback. To better understand the problem, let us consider the controlled CS model, introduced by Caponigro, Fornasier, Piccoli, and Trélat [14, 15]. Consider the control system

$$\begin{cases} \dot{x}_i(t) = v_i(t) \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_j(t) - x_i(t)\|)(v_j(t) - v_i(t)) + u_i(t) \end{cases} \quad i = 1, \dots, N \quad (12)$$

with the bound on the control

$$\sum_{i=1}^N \|u_i(t)\| \leq M \quad (13)$$

for a given $M > 0$. Any $v \in \mathbb{R}^d$ can be written as $v = (\bar{v}, \dots, \bar{v}) + v_\perp$. In [15], the authors proved that the feedback control defined by

$$u(t) = -\alpha v_\perp(t), \quad (14)$$

for $\alpha > 0$, stabilizes the system to flocking (in infinite time) while satisfying condition (13) if α is small enough. Indeed, $\dot{V} \leq -\frac{2}{N} \sum_i \langle v_{\perp i}, u_i \rangle = -2\alpha V$.

Remark 3 The control (14) acts on a large number of agents simultaneously. This is inconvenient for practical purposes, since it requires intensive instantaneous communications between all agents. In what follows, we look at more economical controls that are active on as few components as possible at any instant of time. This leads to the concept of *sparse control*.

5.1.4 Sparse Stabilization

The objectives of sparse stabilization are as follows:

- To design a *sparse feedback control* steering “optimally” the system to flocking, with
 - (i) a minimal amount of components active at each time: concept of *componentwise sparse control*.
 - (ii) a minimal amount of switchings in time: concept of *time sparse control*
- To control the system to any prescribed flocking.

Our idea to promote sparsity is to use ℓ^1 minimization as in image analysis where it has become very popular.

Note that the (far from being sparse) feedback stabilizing control (14) is the solution of the minimization problem

$$\min_{\sum_{i=1}^N \|u_i\| \leq M} \left(\frac{1}{2N^2} \sum_{i,j=1}^N \langle v_i - v_j, u_i - u_j \rangle \right) = \min_{\sum_{i=1}^N \|u_i\| \leq M} \left(\frac{1}{N} \sum_{i=1}^N \langle v_{\perp i}, u_{\perp i} \rangle \right).$$

Instead, we now consider the slightly modified minimization problem

$$\min_{\sum_{i=1}^N \|u_i\| \leq M} \left(\frac{1}{2N^2} \sum_{i,j=1}^N \langle v_i - v_j, u_i - u_j \rangle + \gamma(X) \frac{1}{N} \sum_{i=1}^N \|u_i\| \right),$$

where

$$\gamma(X) = \int_{\sqrt{X}}^{+\infty} a(\sqrt{2N}r) dr.$$

Here, the use of the ℓ^1 norm is to enforce sparsity, and the weight $\gamma(X)$ is used as a threshold implying that the control will switch off when entering the flocking region. The optimal solution of this minimization problem is the *componentwise sparse feedback control* u° defined as

- if $\max_{1 \leq i \leq N} \|v_{\perp i}(t)\| \leq \gamma(X(t))^2$, then $u^\circ(t) = 0$
- if $\|v_{\perp j}(t)\| = \max_{1 \leq i \leq N} \|v_{\perp i}(t)\| > \gamma(X(t))^2$ (with j be the smallest index) then

$$u_j^\circ(t) = -M \frac{v_{\perp j}(t)}{\|v_{\perp j}(t)\|}, \quad \text{and} \quad u_i^\circ(t) = 0 \quad \text{for every } i \neq j.$$

Theorem 1 [14, 15] *The sparse feedback control u° stabilizes the system to flocking.*

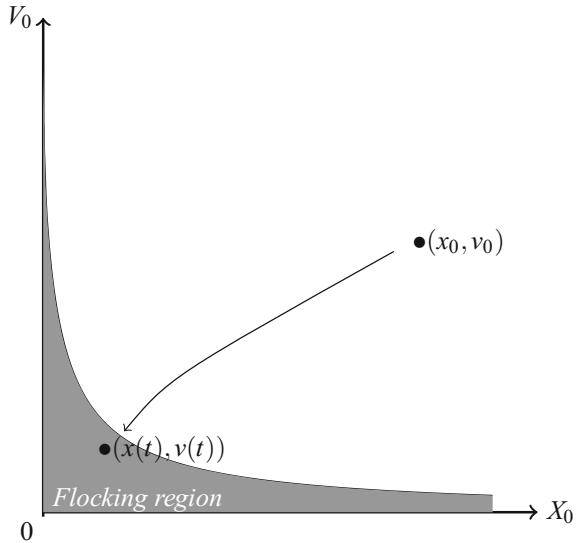
Indeed, we have

$$\dot{V} \leq \frac{2}{N} \sum_i \langle v_{\perp i}, u_i^\circ \rangle = -2 \frac{M}{N} \|v_{\perp j}\|$$

with $\|v_{\perp j}\| = \max_{1 \leq i \leq N} \|v_{\perp i}\| \geq \sqrt{V}$ implying that $\dot{V} \leq -2 \frac{M}{N} \sqrt{V}$, hence any trajectory enters in finite time the flocking region, and then we take $u = 0$ (forever), as illustrated on Figure 9, thus letting the trajectory naturally converge to flocking. Note that, alternatively, one can choose not to switch off the control (even when one has entered the flocking region): In that case, the trajectory reaches flocking within finite time, because $\sqrt{V(t)} \leq \sqrt{V(0)} - 2 \frac{M}{N} t$.

Remark 4 By construction, this feedback control is componentwise sparse. However, it is not necessarily time sparse: It may *chatter*. Indeed, the strategy described above consists of focusing on the agent that is the farthest possible from the mean, in order to steer it closer to the mean. But that agent may change as time evolves and

Fig. 9 Control to flocking.
The flocking region $\sqrt{V_0} \leq \frac{1}{\sqrt{2N}}(\frac{\pi}{2} - \arctan(\sqrt{X_0}))$ corresponds to the CS model (8) with parameter $\beta = 1$, see Proposition 1



oscillations may appear. In order to avoid possible chattering in time, [14, 15] have implemented the classical sample-and-hold procedure [18], consisting of freezing the value of the control over a certain duration, called sampling time. The resulting sampled control is then time sparse, by construction. Therefore, in such a way, we obtain a *time sparse and componentwise sparse feedback control*.

Remark 5 (Sparse is “optimal”). It has been proven in [14, 15] that “sparse is better” in the following sense:

For every time t , $u^\circ(t)$ minimizes $\frac{d}{dt}V(t)$ over all possible feedback controls.

In other words, at every instant of time t , the above feedback control $u^\circ(t)$ is the best choice in terms of the rate of convergence to flocking. This means that a policy maker who is not allowed to have prediction on future developments should always consider more favorable to intervene with stronger actions on the fewest possible instantaneous optimal leaders, rather than trying to control more agents with minor strength.

Remark 6 (Optimal is sparse). The notion of “sparsity” arises naturally in optimal control of multiagent systems. Indeed, consider the following simple linear consensus model with an external control

$$\dot{x}_i = \sum_{j=1^N} a_{ij}(x_j - x_i) + u_i, \quad i = 1, \dots, N,$$

where $x_i \in \mathbb{R}^d$, $a_{ij} \in \mathbb{R}$, and the control $u = (u_1, \dots, u_N)$ verifies the constraint (13) for some given $M > 0$. Consider the problem of steering the system to the consensus $x_1 = \dots = x_N$ in minimal time T . Then, the Pontryagin maximum principle ensures the existence of a absolutely continuous nontrivial vector function $p_1(t), \dots, p_N(t)$ satisfying the adjoint equation

$$\dot{p}_i = \sum_{j=1}^N a_{ij}(p_i - p_j)$$

with final constraint $\sum_i p_i(T) = 0$. The minimality condition for the time optimal control reads

$$\min \sum_{i=1}^N \langle p_i(t), u_i(t) \rangle,$$

so that the optimal control is, if $\|p_i(t)\| > \|p_j(t)\|$ for every $j \neq i$,

$$u_i = -M \frac{p_i(t)}{\|p_i(t)\|} \quad \text{and } u_j = 0 \text{ for } j \neq i.$$

In particular, the time optimal control is sparse except when the index for which $\|p_i(t)\|$ is maximal is not unique. However, in the generic case, in which all interaction coefficients a_{ij} are pairwise distinct, then the time optimal control is sparse for almost every $t \in [0, T]$.

The previous analysis was done on an approachable toy model. In general, finding the optimal strategy for a consensus model can be very hard. However, it is possible to find sparsity features for more complicated optimal control problems, as shown in Section 5.1.6 below.

5.1.5 Sparse Local Controllability Near Consensus

It is possible to show that generically a consensus or an alignment system is controllable near the consensus or alignment manifold by means of a sparse control. More precisely, given a generic pair of configurations sufficiently close to consensus, there exists a strategy, acting on a single agent at every time, that steers the system from one configuration to the other. This property was proven in [14, 15] for alignment systems. The proof relies on the fact that given a generic Laplacian matrix L satisfying some open condition on the coefficients, and any column vector B with only one component different from 0, the linear system

$$\dot{x} = -Lx + Bu$$

verifies the Kalman rank condition for controllability. From this fact, it is possible to infer small-time local controllability for opinion formation models or alignment

models near consensus. Moreover, it is possible to choose the controlled agent a priori. More precisely, the result of [14, 15] is the following:

Proposition 3 *For almost every consensus there exists a neighborhood in which controllability with time sparse and componentwise sparse control holds.*

This result is easy to establish by linearization around a consensus point. The Kalman condition holds for every consensus point verifying and open algebraic condition on the coefficients $a(\|x_i - x_j\|) i, j = 1, \dots, N$ (whence the “almost every” of the statement). First-order models can be dealt with similarly (see also [90] on opinion formation models).

By stabilization and iterated local controllability along a path of consensus points (note that the set of consensus points is arc-connected), we obtain the following result:

Corollary 1 *Any point of $(\mathbb{R}^d)^N \times (\mathbb{R}^d)^N$ can be steered to almost any point of the consensus manifold in finite time by means of a time sparse and componentwise sparse control.*

5.1.6 Optimal Control

In [14, 15], the authors have considered, for the finite-dimensional CS model, the optimal control problem with a fixed initial point and free final point, of minimizing the cost functional

$$\int_0^T \sum_{i=1}^N \left(v_i(t) - \frac{1}{N} \sum_{j=1}^N v_j(t) \right)^2 dt + \gamma \sum_{i=1}^N \int_0^T \|u_i(t)\| dt$$

where $\gamma > 0$ is fixed, under the constraint $\sum_{i=1}^N \|u_i(t)\| \leq M$. As before, the ℓ^1 -norm (with weight γ) implies componentwise sparsity features of the optimal control. The proof is done by applying the Pontryagin maximum principle and by developing genericity arguments.

However, because of the coupling between space and velocity, these properties may not be easy to check in practice. We can note that, if instead of considering the CS model (8), we consider the much simpler HK model (5), then the above optimal control problem becomes quite obvious to analyze. For instance, it is easy to prove that, under generic conditions on the interaction coefficients a_{ij} , the optimal control is componentwise sparse. Such optimal control problems have not yet been considered for the kinetic CS equation (25).

Another interesting example of an optimal control problem involves the *collective migration* model [61, 70], in which the agents (for example, migrating birds) aim to align their velocities to a target migration velocity. In this model, not only do the agents interact with each other to evolve as a group as in the CS model, but they also gather clues from the environment to sense the predetermined migration

velocity V . The control is not an exterior force represented by an additive control as in (12). Instead, the control is considered to reflect an internal decision making process between two possible actions: following the group or sensing the target migration velocity. Each agent balances those two forces via a control function $\alpha_i \in [0, 1]$. The controlled system writes:

$$\begin{cases} \dot{x}_i = v_i \\ \dot{v}_i = \alpha_i(V - v_i) + (1 - \alpha_i)\frac{1}{N}\sum_{j=1}^N a(\|x_j - x_i\|)(v_j - v_i) \end{cases} \quad \text{for } i \in \{1, \dots, N\}, \quad (15)$$

One way to minimize distance from alignment to the target velocity V is to minimize at a given final time the functional $\mathbb{V} = \frac{1}{N}\sum_i \|v_i - V\|^2$ with the constraints $0 \leq \alpha_i \leq 1$ for all $i \in \{1, \dots, N\}$ and $\sum_i \alpha_i \leq M$. The constraint on the total control strength $\sum_i \alpha_i \leq M$ reflects the fact that it is more energy-consuming to sense the target velocity than it is to follow the group, so that only a few individuals can sense it. This naturally divides the group into leaders and followers. Using the Pontryagin Maximum Principle, Piccoli, Pouradier Duteil, and Scharf [70] were able to fully determine an optimal control strategy in the case $M = 1$ and $a \equiv 1$. Interestingly, when the initial average velocity \bar{v} is very close to the target velocity V , the optimal control strategy requires the existence of an initial *inactivation* time, during which no control should be exerted on the system. This is due to the fact that without control, due to its inherent properties, the system naturally relaxes to alignment of all velocities.

5.2 Infinite Dimension

As in Section 2.3.3, we want to consider, in some appropriate way, the mean-field limit of the finite-dimensional controlled CS system (12). A difficulty comes from the fact that, to ensure existence and uniqueness of the resulting kinetic equation, we need minimal regularity properties of the velocity field that do not hold true when considering general controls. Another difficulty is that passing to the limit in the previously designed sparse control makes no sense: Indeed, in the finite-dimensional model, the control has been designed in such a way that, at any instant of time, at most one component of the control is active. But when taking the limit $N \rightarrow +\infty$, this is not feasible and the notion of sparsity must be redefined.

5.2.1 Γ -Limit

A first approach in controlling kinetic equations consists of taking the limit of the finite-dimensional controls, in a sense defined by Fornasier and Solombrino [35]. This combines the concepts of mean-field limit for the probability measure and of

Γ -limit for the control in order to define an appropriate mean-field control for the kinetic equation. More specifically, we study the limit when $N \rightarrow \infty$ of the control problem that consists of finding the minimum of the cost functional

$$\min_f \int_0^T \int_{\mathbb{R}^{2d}} (L(x, v, \mu_N) + \psi(f(t, x, v))) d\mu_N(t, x, v) dt \quad (16)$$

over all control functions f that satisfy:

- (i) $f : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^d$ is a Carathéodory function
- (ii) $f(t, \cdot) \in W_{loc}^{1,\infty}(\mathbb{R}^n, \mathbb{R}^d)$ for almost every $t \in [0, T]$
- (iii) $|f(t, 0)| + \text{Lip}(f(t, \cdot), \mathbb{R}^d) \leq \ell(t)$ for almost every $t \in [0, T]$

where $\ell \in L^q(0, T)$ for a given horizon time $T > 0$ and $1 \leq q < +\infty$. In (16),

$$\mu_N(t, x, v) = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i, v_i)}(x, v)$$

is the atomic measure supported on the phase space trajectories $(x_i(t), v_i(t)) \in \mathbb{R}^{2d}$, constrained by satisfying the system:

$$\begin{cases} \dot{x}_i = v_i, \\ \dot{v}_i = (H \star \mu_N)(x_i, v_i) + f(t, x_i, v_i) \end{cases} \quad (17)$$

with initial datum $\mu_N^0(t, x, v) = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i^0, v_i^0)}(x, v)$. The notation $H \star \mu_N$ denotes the convolution of H with μ_N , where $H : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a sublinear and locally Lipschitz continuous interaction kernel. The functions L and ψ are taken to satisfy appropriate conditions [35]. Applications of this problem include finding ways to influence large crowds, for instance to guide them through an exit in emergency situations. In this context, the function f represents an external control on the crowd [34].

If there exists a compactly supported limit μ^0 to the sequence of atomic measures μ_N when the number of agents N tends to infinity in the sense of the Wasserstein distance (i.e., $\lim_{N \rightarrow \infty} \mathcal{W}_1(\mu_N^0, \mu^0) = 0$), then there exists a subsequence $(f_{N_k})_{k \in \mathbb{N}}$ and a function f_∞ such that f_{N_k} Γ -converges to f_∞ and f_∞ is a solution of the infinite-dimensional optimal control problem

$$\min_f \int_0^T \int_{\mathbb{R}^{2d}} (L(x, v, \mu) + \psi(f(t, x, v))) d\mu(t, x, v) dt \quad (18)$$

where μ is the unique weak solution of the kinetic equation

$$\frac{\partial \mu}{\partial t} + \langle v, \nabla_x \mu \rangle = \operatorname{div}_v ((H \star \mu + f) \mu) \quad (19)$$

with initial datum μ^0 .

5.2.2 Control by Leaders

A common way to control a large crowd is to act on a selected few individuals that will behave as leaders to guide the crowd. In the case of finite-dimensional systems, it is frequent to look for controls that are vanishing for most of the agents and for most of the time. These strategies are referred to as sparse control, as seen in Section 5.1. Such controls have obvious advantages, being both moderate in the external control and parsimonious in the number of agents controlled. Extending the concept of leaders to mean-field limits is not straightforward. Indeed, when representing the crowd with a particle density distribution, the action of a finite number of agents becomes negligible compared to the size of the crowd. Albi and Pareschi have looked at the microscopic-macroscopic limit of such systems, see [4]. In [34], Fornasier, Piccoli, and Rossi solve this problem by using a mixed granular-diffuse description of the crowd and prove convergence of the solution of the finite-dimensional problem when the number of followers tends to infinity to the solution of this new system. Similar approaches involving the coupling of microscopic dynamics for the leaders and macroscopic dynamics for the followers were adopted by Albi, Bongini, Cristiani, and Kalise in [1] and by Colombo and Pogodaev in [19]. Here, we present the approach of [34]. Let (y_k, w_k) denote the space-velocity variables of the m leaders of the crowd, and (x_i, v_i) those of the N followers, so that for a given locally Lipschitz interaction kernel with sublinear growth H ,

$$\begin{cases} \dot{y}_k = w_k \\ \dot{w}_k = H \star \mu_N(y_k, w_k) + H \star \mu_m(y_k, w_k) + u_k & \text{for } k \in \{1, \dots, m\} \\ \dot{x}_i = v_i \\ \dot{v}_i = H \star \mu_N(x_i, v_i) + H \star \mu_m(x_i, v_i) & \text{for } i \in \{1, \dots, N\} \end{cases} \quad (20)$$

where

$$\mu_N(t, x, v) = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i(t), v_i(t))} \text{ and } \mu_m(t, y, w) = \frac{1}{m} \sum_{k=1}^m \delta_{(y_k(t), w_k(t))} \quad (21)$$

and $u_k : [0, T] \rightarrow \mathbb{R}^d$ are measurable controls for $k \in \{1, \dots, m\}$. The optimal control problem consists of finding solutions of

$$\min_{u_k} \int_0^T (L(y(t), w(t), \mu_N(t)) + \frac{1}{m} \sum_{k=1}^m |u_k(t)|) dt. \quad (22)$$

where $(y(t), w(t), \mu_N(t))$ are subject to the dynamics (20).

The authors of [34] showed that a mean-field limit of system (20) when N tends to infinity can be derived as the coupling of controlled ODEs for the evolution of the leaders' positions and velocities and of a PDE for the compactly supported probability measure μ of the followers in the position-velocity space:

$$\begin{cases} \dot{y}_k = w_k \\ \dot{w}_k = H \star (\mu + \mu_m)(y_k, w_k) + u_k \quad \text{for } k \in \{1, \dots, m\} \\ \partial_t \mu + \langle v, \nabla_x \mu \rangle = \operatorname{div}_v((H \star (\mu + \mu_m))\mu). \end{cases} \quad (23)$$

Moreover, the optimal controls u_N^* of the finite-dimensional control problem (22)–(20) converge weakly for $N \rightarrow \infty$ to optimal controls u^* that are solutions of

$$\min_{u_k} \int_0^T (L(y(t), w(t), \mu(t)) + \frac{1}{m} \sum_{k=1}^m |u_k(t)|) dt. \quad (24)$$

where $(y(t), w(t), \mu(t))$ are subject to the dynamics (23). In [33], these results were extended to Mayer-type minimization problems. Such formulations are particularly useful in that they combat the *curse of dimensionality*. Indeed, even though the total number of agents is allowed to tend to infinity, the number of controlled ones stays bounded and small, which keeps the numerical computations feasible.

5.2.3 Controlled Kinetic Cucker–Smale Model

The first two approaches to controlling kinetic systems reported in Sections 5.2.1 and 5.2.2 consist of finding an optimal control for the finite-dimensional system and passing to the limit (in some appropriate sense) when the number of agents tends to infinity. Another approach involves controlling the PDE directly. This was done, for instance, by Piccoli, Rossi, and Trélat [71]. A difficulty arises when the control is componentwise sparse (see Section 5.1.4). Keeping just one component of the control active is only practical in finite dimension. One way to translate this criterion to infinite-dimensional problems is to pass to the limit by keeping proportions: Given some fixed $c \in (0, 1)$, assume that the control acts on cN agents of the N -sized group. It is then easy to see how to modify the sparse control designed for the finite-dimensional model in order to fit this new requirement, and it makes sense to pass to the limit $N \rightarrow +\infty$. The real number c represents the proportion of the crowd on which one is allowed to act. When passing to the limit, we obtain a control domain, denoted $\omega(t)$ in what follows, representing the controlled part of the crowd at timebreak t .

Let us now be more precise. Following [71], we consider the mean-field limit

$$\partial_t \mu + \langle v, \nabla_x \mu \rangle + \operatorname{div}_v ((\xi[\mu] + \chi_\omega u) \mu) = 0 \quad (25)$$

with $u \in L^\infty(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d)$ and $\omega(t) \subset \mathbb{R}^d$ measurable, such that

$$\|u(t, \cdot, \cdot)\|_{L^\infty(\mathbb{R}^d \times \mathbb{R}^d)} \leq 1, \quad (26)$$

standing for a bounded external action, and

$$\mu(t)(\omega(t)) = \int_{\omega(t)} d\mu(t)(x, v) \leq c, \quad (27)$$

modeling that one is allowed to act only on *a given proportion c of the crowd*. This is the natural notion of *sparse control* in the infinite-dimensional setting.

A possible variant is to consider the constraint

$$|\omega(t)| = \int_{\omega(t)} dx dv \leq c, \quad (28)$$

representing a limit on the space of configurations.

Hence, now the control is $\chi_\omega u$ and consists of choosing, at any instant of time, the control domain $\omega(t)$, and the force $u(t, \cdot, \cdot)$ with which one acts on the crowd along the control domain. Note that it is not so common in the existing literature to control not only an external force but also a domain.

Theorem 2 [71] *For every $\mu^0 \in \mathcal{P}_c^{ac}(\mathbb{R}^d \times \mathbb{R}^d)$, there exists a control $\chi_\omega u$ satisfying (26) and either (27) or (28), and the corresponding solution $\mu \in C^0(\mathbb{R}; \mathcal{P}_c^{ac}(\mathbb{R}^d \times \mathbb{R}^d))$ such that $\mu(0) = \mu^0$ converges to consensus.*

Remark 7 The strategy to prove this theorem is quite long and technical, and is not reported in detail here. We just give hereafter the main intuitive idea. Writing the controlled kinetic CS equation (25) as

$$\partial_t \mu + \operatorname{div}_{(x,v)} (V_{\omega,u}[\mu] \mu) = 0, \quad (29)$$

with the controlled velocity field

$$V_{\omega,u}[\mu] = \begin{pmatrix} v \\ \xi[\mu] + \chi_\omega u \end{pmatrix},$$

the *controlled particle flow* $\Phi_{\omega,u}(t)$ generated by $V_{\omega,u}[\mu(t)]$ yields the characteristics

$$\dot{x}(t) = v(t), \quad \dot{v}(t) = \xi[\mu(t)](x(t), v(t)) + \chi_{\omega(t)} u(t, x(t), v(t)).$$

This is a control system, describing the (controlled) motion of particles. As in the uncontrolled case, the measure, solution of the kinetic equation, is then the pushforward of the initial measure:

$$\mu(t) = \Phi_{\omega,u}(t)\#\mu^0.$$

Having these facts in mind, we adopt, as in the finite-dimensional case, a *shepherd control design strategy*: At every instant of time, we choose $\omega(t)$ and $u(t)$ such that the controlled velocity field $V_{\omega,u}[\mu(t)]$ points inward the invariant domain, thus confining the population. This implies that the size of $\text{supp}_v(\mu(t))$ (velocity support of the measure) decreases exponentially in time. This construction is carried out in [71], piecewise in time, and in an algorithmic way, thus resulting in an explicit control strategy such that

- $\omega(t)$ is piecewise constant in t ,
- $u(t, x, v)$ is piecewise constant in t for (x, v) fixed, C^0 and piecewise linear in (x, v) for t fixed.

An important difficulty in dealing with the kinetic equation (29) is to ensure existence and uniqueness of the solution. Indeed, for general controls, the velocity field is not regular enough. An essential feature of the control strategy designed in [71] is that the control is piecewise smooth, and then existence and uniqueness of the solution are ensured in an iterative way. Actually, the solution μ of (29) remains absolutely continuous and of compact support, and it becomes singular only in infinite time.

Note that, as in the finite-dimensional setting, the control is switched off when entering the consensus region: Given any $\mu^0 \in \mathcal{P}_c^{ac}(\mathbb{R}^d \times \mathbb{R}^d)$, there exists $T(\mu^0) \geq 0$ such that $u(t, x, v) = 0$ for every $t > T(\mu^0)$. Then, the solution reaches consensus (here, a Dirac mass) in infinite time and remains absolutely continuous in between.

Another variant is not to impose a constraint on the control but to penalize its spread in the cost function, as done in [35]. More specifically, for a PDE constrained problem

$$\partial_t \mu + \langle v, \nabla_x \mu \rangle = \text{div}_v((H \star \mu + f)\mu), \quad (30)$$

the control f is the minimizer of the chosen cost:

$$\mathcal{E}_\psi(f) := \int_0^T \int_{\mathbb{R}^{2d}} (L(x, v, \mu) + \psi(f(t, x, v))) d\mu(t, x, v) dt \quad (31)$$

where a relevant choice for ψ is, for instance, $\psi(\cdot) := \gamma |\cdot|$ for $\gamma > 0$. This promotes the sparsity of f thanks to the ℓ^1 norm penalization.

5.2.4 Boltzmann-Type Control for Consensus Dynamics

As stressed in the previous sections, controlling a large number of agents is computationally expensive, and even sometimes unfeasible. For instance, the Pontryagin

maximum principle applied to the minimization problem

$$\min_{u(t) \in [u_L, u_R]} \int_0^T \frac{1}{N} \sum_{j=1}^N \left(\frac{1}{2} (x_j - x_d)^2 + \frac{\kappa}{2} u^2 \right) ds, \quad (32)$$

where x_d is the desired state, for the controlled system

$$\dot{x}_i = \frac{1}{N} \sum_{j=1}^N a(x_i, x_j)(x_j - x_i) + u \quad (33)$$

requires solving the equation for the adjoint vector backward in time over the whole interval $[0, T]$, which is extremely costly for a large number of agents. In [2], Albi, Herty, and Pareschi develop an alternative approach consisting of solving the control problem on a sequence of reduced time intervals. This iterative method is called *model predictive control*. The horizon-receding strategy allows to embed the minimization of the cost functional into the particle interactions.

As done in [2], let $I = [-1, 1]$ represent a bounded set of opinions such that $x_i(t) \in I$, $i = 1, \dots, N$. Alternatively, in the multidimensional setting, one can take $I = \mathbb{S}^d$. Consider the case of binary Boltzmann dynamics with two interacting agents i and j , then their positions x^{n+1} at time $(n+1)\Delta t$ depend on the previous state in the following way:

$$\begin{cases} x_i^{n+1} = x_i^n + \frac{\Delta t}{2} a(x_i^n, x_j^n)(x_j^n - x_i^n) + \frac{\Delta t}{2} U(x_i^n, x_j^n) \\ x_j^{n+1} = x_j^n + \frac{\Delta t}{2} a(x_i^n, x_j^n)(x_j^n - x_i^n) + \frac{\Delta t}{2} U(x_j^n, x_i^n) \end{cases}. \quad (34)$$

The model predictive control performed on a single prediction time horizon allows the explicit expression:

$$\eta U(x_i^n, x_j^n) = \frac{\beta}{2} \left((x_d - x_j^n) + (x_d - x_i^n) + \eta(a(x_i, x_j) - a(x_j, x_i))(x_j^n - x_i^n) \right) \quad (35)$$

where $\beta := \frac{2\eta}{\kappa+2\eta}$ and $\eta = \Delta t/2$. The kinetic Boltzmann equation is obtained by introducing the density distribution of particles $\mu(x, t)$ belonging to the space of probability measures. Then, two agents x and y modify their states according to

$$\begin{cases} x^* = x + \eta(a(x, y)(y - x) + U(x, y)) \\ y^* = y + \eta(a(y, x)(x - y) + U(y, x)) \end{cases} \quad (36)$$

For a test function $\phi(x)$, we write:

$$\frac{d}{dt} \int_I \phi(x) \mu(x, t) dx = \lambda \int_{I^2} (\phi(x^*) + \phi(y^*) - \phi(x) - \phi(y)) \mu(x, t) \mu(y, t) dxdy \quad (37)$$

where λ represents a constant rate of interaction and we considered that $a(x, y) = a(y, x)$. This allows us to show that the limit of the average position $m_\infty := \lim_{t \rightarrow \infty} m(t)$ where $m(t) = \int_I x \mu(x, t) dx$ stays close to the desired state x_d , and in the symmetric case $a(x, y) = a(x, y)$, we even have $m_\infty = x_d$. Moreover, if the interaction kernel is simplified to $a(x, y) = 1$, then one shows that the particle distribution converges to the Dirac measure $\delta(x - x_d)$ centered in the desired state x_d , which implies that the system reaches consensus.

To derive the asymptotic limit of the model while retaining the memory of the binary interactions (36), one can refer to the so-called *quasi-invariant opinion limit* [84]. This consists of adapting to the context of consensus models the concept of *grazing collision limit* used to consider longtime solutions of the Boltzmann equation [87]. For a summary of these concepts, see [67]. Here, this is done by rescaling time in (37). In particular, taking $\eta = \varepsilon$, $\lambda = 1/\varepsilon$, the limit when ε tends to zero leads to a controlled kinetic equation of type

$$\mu_t = \operatorname{div}_x((\xi[\mu] + \zeta[\mu])\mu),$$

where

$$\xi[\mu](x) = \int_I a(x, y)(y - x)\mu(y)dy$$

and

$$\zeta[\mu](x) = \int_{\mathbb{R}^d} K(x, y)d\mu(y), \text{ for } K(x, y) = \frac{1}{\kappa}((x_d - x) + (x_d - y)).$$

This approach can be easily extended to the case of the CS model and leader–follower model, see [3, 5].

5.2.5 Mean-Field Games

Another common way to deal with control of large systems is to use *mean-field games*, a theory that was introduced by Lasry and Lions in 2006 [38, 59] and independently by Caines [13, 48]. A wealth of results have since then been obtained in this game-theoretic setting, considering that each agent makes a decision in order to optimize a given cost based on its available information (see Degond, Liu and Ringhofer [27]). For instance, in applications to economics, it is meaningful to study Nash equilibria, a stable state in which no agent can improve its cost by changing alone its strategy.

Mean-field games are used to consider each agent's individual decision, given his knowledge of the system. Consequently, most applications of mean-field games are found in economics, where each agent strives to optimize its wealth given the current state of the market (Guéant, Lasry and Lions [38]). For instance, price formation models can be derived by dividing the population into buyers and sellers. Other applications involve crowd modeling (Lachapelle and Wolfram [58]).

In [7, 8], Bardi and Priuli derive the mean-field limit of a stochastic differential game with N players:

$$dX_t^i = (A^i X_t^i - \alpha_t^i) dt + \sigma^i dW_t^i, \quad X_0^i \in \mathbb{R}^d, \quad i = 1, \dots, N. \quad (38)$$

where A^i and σ^i are $d \times d$ matrices, $(W_t^i)_{i \in \{1, \dots, N\}}$ are N independent d -dimensional standard Brownian motions, and α_t^i is a process adapted to W_t^i representing the control of the i -th player, designed to minimize the quadratic running cost

$$J^i(X, \alpha^1, \dots, \alpha^N) := \liminf_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\int_0^T \frac{(\alpha_t^i)^T R^i \alpha_t^i}{2} + (X_t - \bar{X}_i)^T Q^i (X_t - \bar{X}_i) dt \right]. \quad (39)$$

Here, \mathbb{E} denotes the expected value, R^i are positive definite $d \times d$ matrices, Q^i are symmetric $Nd \times Nd$ matrices, and \bar{X}_i are reference positions. Following the approach of [59], the corresponding system of N nonlinear Hamilton–Jacobi–Bellman PDEs coupled with N Kolmogorov–Fokker–Planck equations can be derived as follows:

$$\begin{cases} -\text{tr}(v^i D^2 v^i) + \mathcal{H}^i(x, \nabla v^i) + \lambda^i = f^i(x; m^1, \dots, m^N) \\ -\text{tr}((v^i D^2 m^i) - \text{div}(m^i \frac{\partial \mathcal{H}^i}{\partial p}(x, \nabla v^i))) = 0 \\ \int_{\mathbb{R}^d} m^i(x) dx = 1, \quad m^i > 0, \end{cases} \quad (40)$$

where the unknowns are the functions v^i , the scalars λ^i and the measures m^i , and \mathcal{H}^i denotes the i -th Hamiltonian. This leads to Nash equilibria obtained by affine feedback. In order to derive the mean-field limit of (38), we need to consider that the players are *nearly identical*; that is, that they are influenced in the same way by pairs of other players, have the same control systems ($A^i = A$), the same costs of the controls ($R^i = R$), the same reference positions ($\bar{X}_i = X_d$), and the same primary costs of displacement. Then, given suitable conditions, there exist unique solutions to the system of HJB-KFP equations (40). Moreover, those solutions converge when $N \rightarrow \infty$ to the solutions of a mean-field system of the form:

$$\begin{cases} -\text{tr}(v D^2 v) + \nabla v^T \frac{R^{-1}}{2} \nabla v - \nabla v^T A x + \lambda = \hat{V}[m](x) \\ -\text{tr}((v D^2 m) - \text{div}(m \cdot R^{-1} \nabla v - A x)) = 0 \\ \int_{\mathbb{R}^d} m(x) dx = 1, \quad m > 0. \end{cases} \quad (41)$$

Mean-field games consist of optimizing control strategies over a large time horizon. This is both computationally expensive and not always realistic, since individual agents might not have access to information on the state of the system in the distant future. Instead, some strategies called *best-reply strategies* compute the best instantaneous response given the present state of the system, by steepest gradient descent [27]. This type of control is suboptimal in the long term, but is in some ways more realistic and more feasible. A good compromise between standard large-time mean-

field approaches and best-reply strategies consists of minimizing the cost function over a small shifting time horizon [2], as done in *Model Predictive Control*, see Maciejowski, Goulart, and Kerrigan [62] and Degond, Herty, and Liu [26].

6 Generalizations

There exist many possible generalizations of the models considered above. One way to better adapt the model to the phenomenon of interest is to use general interaction potentials. For example, in the case of animal group modeling, Carillo et al propose to take into account the cone of vision of each animal i to define its influential set of neighbors \mathcal{N}_i [17]. This naturally singles out a certain number of instantaneous leaders, defined as the animals whose cones of vision are pointed outward so that they do not follow any other agent. Such dynamics are expected to lead to clustering of the group into a finite number of subgroups each following a leader.

In [21], Cristiani, Frasca, and Piccoli studied the effect of anisotropic interaction regions on the shape of the group. Around each agent are defined a zone of attraction and a zone of repulsion that can each be isotropic or anisotropic. Depending on the nature of those interactions, simulations show that various patterns can be obtained in the group: crystal-like clusters of individuals, lines, or V-like formations.

In [32], D’Orsogna, Chuang, Bertozzi, and Chayes propose a model to take into account self-propelling, friction, and attraction–repulsion effects. More specifically, each agent’s velocity is defined as $\dot{v}_i = (\alpha - \beta \|v_i\|^2)v_i - \nabla_i U(x_i)$, where $U(x_i)$ is the Morse potential used to include attractive and repulsive ranges. For agent i , $U(x_i) = \sum_{j \neq i} (C_r e^{-\|x_i - x_j\|/l_r} + C_a e^{-\|x_i - x_j\|/l_a})$ where l_r and l_a are, respectively, the repulsive and attractive ranges and C_r and C_a the repulsive and attractive amplitudes. In the case of animal groups or other biological applications, the most relevant cases are $l_a > l_r$ and $C_a > C_r$, for short-range repulsion and long-range attraction. The parameter α models the self-propulsion capacity of agent i , while the parameter β represents a friction according to Rayleigh’s law. This model gives rise to various types of regimes, depending on the ratios l_r/l_a and C_r/C_a . The system is said to be *H*-stable if the total potential energy is bounded below by a multiple of the number of agents N , i.e., $U \geq -BN$ for some constant $B \geq 0$. *H*-stability ensures that the system does not collapse when $N \rightarrow \infty$, so that the particles form a crystal-like structure. When the system is not *H*-stable, it is said to be *catastrophic*, and as $N \rightarrow \infty$, the inter-particle intervals shrink to zero.

Another common generalization of the models described in this chapter consists of adding white noise to the dynamics. For example, in [91], Yates et al. argue that locusts use white noise to maintain swarm alignment. This claim is supported by experimental evidence itself modeled using the kinetic Fokker–Planck equation with noise. See also [40] for stochastic models.

Acknowledgements The authors acknowledge the partial support of the NSF Project “KI-Net,” DMS Grant # 1107444.

References

1. G. Albi, M. Bongini, E. Cristiani, and D. Kalise. Invisible control of self-organizing agents leaving unknown environments. *SIAM Journal on Applied Mathematics*. to appear.
2. G. Albi, M. Herty, and L. Pareschi. Kinetic description of optimal control problems and applications to opinion consensus. *Communications in Mathematical Sciences*, 13(6):1407–1429, 2015.
3. G. Albi and L. Pareschi. Selective model-predictive control for flocking systems. preprint.
4. G. Albi and L. Pareschi. Modeling of self-organized systems interacting with a few individuals: from microscopic to macroscopic dynamics. *Applied Mathematics Letters*, 26:397–401, 2013.
5. G. Albi, L. Pareschi, and M. Zanella. Boltzmann-type control of opinion consensus through leaders. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 372(2028), 2014.
6. 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.
7. M. Bardi and F. S. Priuli. LQG mean-field games with ergodic cost. In *52nd IEEE Conference on Decision and Control*, pages 2493–2498, Dec 2013.
8. M. Bardi and F. S. Priuli. Linear-quadratic n -person and mean-field games with ergodic cost. *SIAM Journal on Control and Optimization*, 52(5):3022–3052, 2014.
9. L. Behera and F. Schweitzer. On spatial consensus formation: Is the Sznajd model different from a voter model? *International Journal of Modern Physics C*, 14(10):1331–1354, 2003.
10. V. D. Blondel, J. M. Hendrickx, and J. N. Tsitsiklis. Continuous-time average-preserving opinion dynamics with opinion-dependent communications. *SIAM Journal on Control and Optimization*, 48(8):5214–5240, 2010.
11. F. Bullo, J. Cortés, and S. Martínez. Distributed control of robotic networks: a mathematical approach to motion coordination algorithms. *Princeton series in applied mathematics. Princeton University Press, Princeton*, 2009.
12. J. A. Cañizo, J. A. Carrillo, and J. Rosado. A well-posedness theory in measures for some kinetic models of collective motion. *Mathematical Models and Methods in Applied Sciences*, 21(03):515–539, 2011.
13. P. E. Caines. *Encyclopedia of Systems and Control*, chapter Mean Field Games, pages 1–6. Springer London, London, 2013.
14. M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and optimal control of the Cucker–Smale model. *Mathematical Control and Related Fields*, 3:447–466, 2013.
15. M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and control of alignment models. *Mathematical Models and Methods in Applied Sciences*, 25(3):521–564, 2015.
16. M. Caponigro, A. C. Lai, and B. Piccoli. A nonlinear model of opinion formation on the sphere. *Discrete and Continuous Dynamical Systems Ser. A*, (9):4241–4268, 2015.
17. J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, chapter Particle, kinetic, and hydrodynamic models of swarming, pages 297–336. Birkhäuser Boston, Boston, 2010.
18. F. H. Clarke, Y. S. Ledyaev, E. D. Sontag, and A. I. Subbotin. Asymptotic controllability implies feedback stabilization. *Automatic Control, IEEE Transactions on*, 42(10):1394–1407, 1997.
19. R. Colombo and N. Pogodaev. On the control of moving sets: positive and negative confinement results. *SIAM J. Control Optim.*, 51(1):380–401, 2013.
20. I. 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.
21. E. Cristiani, P. Frasca, and B. Piccoli. Effects of anisotropic interactions on the structure of animal groups. *Journal of mathematical biology*, 62(4):569–588, 2011.
22. E. Cristiani, B. Piccoli, and C. Tosin. Multiscale modeling of granular flows with application to crowd dynamics. *SIAM Multiscale Modeling and Simulations*, 9:155–182, 2011.

23. F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52:852–862, 2007.
24. M. H. De Groot. Reaching a consensus. *Journal of American Statistical Association*, 69:118 – 121, 1974.
25. G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.
26. P. Degond, M. Herty, and J.-G. Liu. Mean-field games and model predictive control. preprint.
27. P. Degond, J.-G. Liu, and C. Ringhofer. Large-scale dynamics of mean-field games driven by local Nash equilibria. *Journal of Nonlinear Science*, 24(1):93–115, 2013.
28. P. Degond and S. Motsch. Continuum limit of self-driven particles with orientation interaction. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1193–1215, 2008.
29. P. Degond and S. Motsch. Large scale dynamics of the persistent turning walker model of fish behavior. *Journal of Statistical Physics*, 131(6):989–1021, 2008.
30. J. C. Dittmer. Diskrete nichtlineare modelle der konsensbildung. *Diploma thesis Universität Bremen*, 2000.
31. F. Dörfler, M. Chertkov, and F. Bullo. Synchronization in complex oscillator networks and smart grids. *Proceedings of the National Academy of Sciences*, 110(6):2005–2010, 2013.
32. M. R. D’Orsogna, Y. L. Chuang, A. L. Bertozzi, and L. S. Chayes. Self-propelled particles with soft-core interactions: Patterns, stability, and collapse. *Phys. Rev. Lett.*, 96:104302, Mar 2006.
33. M. Fornasier, B. Piccoli, N. Pouradier Duteil, and F. Rossi. Mean-field optimal control by leaders. In *53rd IEEE Conference on Decision and Control*, pages 6957–6962, Dec 2014.
34. M. Fornasier, B. Piccoli, and F. Rossi. Mean-field sparse optimal control. *Philosophical Transaction of the Royal Society A*, 372, 2014.
35. M. Fornasier and F. Solombrino. Mean-field optimal control. *ESAIM: Control, Optimisation and Calculus of Variations*, 20(4):1123–1152, 2014.
36. J. R. P. French. A formal theory of social power. *Psychological Review*, 63:181–194, 1956.
37. I. Giardina. Collective behavior in animal groups: theoretical models and empirical studies. *Human Frontier Science Program Journal*, (205–219), 2008.
38. O. Guéant, J.-M. Lasry, and P.-L. Lions. *Paris-Princeton Lectures on Mathematical Finance 2010*, chapter Mean Field Games and Applications, pages 205–266. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
39. S. Y. Ha, T. Ha, and J. H. Kim. Emergent behavior of a Cucker–Smale type particle model with nonlinear velocity couplings. *IEEE Transactions on Automatic Control*, 55(7):1679–1683, July 2010.
40. S.-Y. Ha, K. Lee, and D. Levy. Emergence of time-asymptotic flocking in a stochastic Cucker–Smale system. *Commun. Math. Sci.*, 7(2):453–469, 06 2009.
41. S.-Y. Ha and E. Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. arXiv preprint [arXiv:0806.2182](https://arxiv.org/abs/0806.2182), 2008.
42. F. Harary. A criterion for unanimity in french’s theory of social power. *Cartwright D (Ed.), Studies in Social Power*, 1959.
43. J. Haskovec. Flocking dynamics and mean-field limit in the Cucker–Smale-type model with topological interactions. *Physica D: Nonlinear Phenomena*, 261:42 – 51, 2013.
44. R. Hegselmann and A. Flache. Understanding complex social dynamics – a plea for cellular automata based modelling. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998.
45. R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.
46. M. Herty, L. Pareschi, and S. Steffensen. Mean–field control and Riccati equations. *Networks and Heterogeneous Media*, 10(3):699–715, 2015.
47. J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
48. M. Huang, R. P. Malham, and P. E. Caines. Large population stochastic dynamic games: closed-loop McKean–Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.*, 6(3):221–252, 2006.

49. A. Huth and C. Wissel. The simulation of the movement of fish schools. *Journal of Theoretical Biology*, 156:365–385, 1992.
50. A. Isidori. *Nonlinear control systems*. Springer Science & Business Media, 2013.
51. P. Jabin and S. Motsch. Clustering and asymptotic behavior in opinion formation. *Journal of Differential Equations*, 257(1):4165–4187, 12 2014.
52. A. Jadbabaie, J. Lin, and A. S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *Automatic Control, IEEE Transactions on*, 48(6):988–1001, 2003.
53. J. M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 08 2000.
54. J. Krause and G. Ruxton. Living in groups. *Oxford series in ecology and evolution*. Oxford University Press, New York, 2002.
55. U. Krause. Soziale dynamiken mit vielen interakteuren, eine problemskizze. *Krause U and Stöckler M (Eds.) Modellierung und Simulation von Dynamiken mit vielen interagierenden Akteuren*, Universität Bremen, pages 37 – 51, 1997.
56. U. Krause. A discrete nonlinear and non—autonomous model of consensus formation. *Elaydi S, Ladas G, Popenda J and Rakowski J (Eds.), Communications in Difference Equations*, Amsterdam: Gordon and Breach Publ., pages 227 – 236, 2000.
57. Y. Kuramoto. Cooperative dynamics of oscillator community a study based on lattice of rings. *Progress of Theoretical Physics Supplement*, 79:223–240, 1984.
58. A. Lachapelle and M.-T. Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation Research Part B: Methodological*, 45(10):1572–1589, 2011.
59. J.-M. Lasry and P.-L. Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.
60. K. Lehrer. Social consensus and rational agnoiology. *Synthese*, 31:141 – 160, 1975.
61. N. Leonard. Multi-agent system dynamics: Bifurcation and behavior of animal groups. *Plenary paper IFAC Symposium on Nonlinear Control Systems, Toulouse, France.*, 2013.
62. J. Maciejowski, P. Goulart, and E. Kerrigan. *Constrained Control Using Model Predictive Control*, pages 273–291. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.
63. L. Moreau. Stability of continuous-time distributed consensus algorithms. In *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, volume 4, pages 3998–4003. IEEE, 2004.
64. L. Moreau. Stability of multiagent systems with time-dependent communication links. *Automatic Control, IEEE Transactions on*, 50(2):169–182, 2005.
65. S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.
66. R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.
67. L. Pareschi and G. Toscani. *Interacting multiagent systems: kinetic equations and Monte Carlo methods*. OUP Oxford, 2013.
68. J. Parrish, S. Viscido, and D. Grunbaum. Self-organized fish schools: an examination of emergent properties. *The Biological Bulletin*, 202:296–305, 2002.
69. L. Perea, P. Elosegui, and G. Gómez. Extension of the Cucker–Smale control law to space flight formations. *Journal of Guidance, Control, and Dynamics*, 32:527–537, 2009.
70. B. Piccoli, N. Pouradier Duteil, and B. Scharf. Optimal control of a collective migration model. *Mathematical Models and Methods in Applied Sciences (to appear)*, 2015.
71. B. Piccoli, F. Rossi, and E. Trélat. Control to flocking of the kinetic Cucker–Smale model. *SIAM Journal on Mathematical Analysis*, 47(6):4685–4719, 2015.
72. A. Rahmani, M. Ji, M. Mesbahi, and M. Egerstedt. Controllability of multi-agent systems from a graph-theoretic perspective. *SIAM Journal on Control and Optimization*, 48(1):162–186, 2009.
73. A. Sarlette. *Geometry and symmetries in coordination control*. PhD thesis, Université de Liège, 2009.
74. A. Sarlette and R. Sepulchre. Consensus optimization on manifolds. *SIAM Journal on Control and Optimization*, 48(1):56–76, 2009.
75. L. Scardovi, A. Sarlette, and R. Sepulchre. Synchronization and balancing on the N-torus. *Systems & Control Letters*, 56(5):335 – 341, 2007.

76. R. Sepulchre. Consensus on nonlinear spaces. *Annual reviews in control*, 35(1):56–64, 2011.
77. R. Sepulchre, D. Paley, N. E. Leonard, et al. Stabilization of planar collective motion: All-to-all communication. *Automatic Control, IEEE Transactions on*, 52(5):811–824, 2007.
78. R. Sepulchre, D. Paley, N. E. Leonard, et al. Stabilization of planar collective motion with limited communication. *Automatic Control, IEEE Transactions on*, 53(3):706–719, 2008.
79. P. Sobkowicz. Modelling opinion formation with physics tools: Call for closer link with reality. *Journal of Artificial Societies and Social Simulation*, 12(1):11, 2009.
80. E. D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.
81. S. H. Strogatz. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1–4):1 – 20, 2000.
82. D. Sumpter. The principles of collective animal behaviour. *Philosophical Transaction of the Royal Society B*, 361:5–22, 2006.
83. K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.
84. G. Toscani. Kinetic models of opinion formation. *Commun. Math. Sci.*, 4(3):481–496, 09 2006.
85. J. N. Tsitsiklis. *Problems in Decentralized Decision making and Computation*. PhD thesis, MIT, 1984.
86. 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.*, 75:1226–1229, Aug 1995.
87. C. Villani. On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations. *Archive for Rational Mechanics and Analysis*, 143(3):273–307, 1998.
88. D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 06 1998.
89. H. Whitney. On singularities of mappings of Euclidean spaces. I. mappings of the plane into the plane. *Annals of Mathematics*, 62(3):374–410, 1955.
90. S. Wongkaew, M. Caponigro, and A. Borzi. On the control through leadership of the Hegselmann-Krause opinion formation model. *Mathematical Models and Methods in Applied Sciences*, 25(03):565–585, 2015.
91. 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*, 106(14):5464–5469, 2009.

Variational Mean Field Games

Jean-David Benamou, Guillaume Carlier and Filippo Santambrogio

Abstract This paper is a brief presentation of those mean field games with congestion penalization which have a variational structure, starting from the deterministic dynamical framework. The stochastic framework (i.e., with diffusion) is also presented in both the stationary and dynamic cases. The variational problems relevant to MFG are described via Eulerian and Lagrangian languages, and the connection with equilibria is explained by means of convex duality and of optimality conditions. The convex structure of the problem also allows for efficient numerical treatment, based on augmented Lagrangian algorithms, and some new simulations are shown at the end of the paper.

1 Introduction and Modeling

The theory of mean field games has been introduced some years ago by Lasry and Lions (in [22–24]) to describe the evolution of a population, where each agent has to choose a strategy, in the form of a trajectory in a state space, which best fits his preferences, but is affected by the other agents through a global mean field effect (with a terminology borrowed from physics).

J.-D. Benamou
INRIA-MOKAPLAN, 2 rue Simone Iff, 75012 Paris, France
e-mail: jean-david.benamou@inria.fr

G. Carlier
Ceremade Univ. Paris Dauphine and INRIA-MOKAPLAN, Paris, France
e-mail: carlier@ceremade.dauphine.fr

F. Santambrogio (✉)
Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay,
91405 Orsay Cedex, France
e-mail: filippo.santambrogio@math.u-psud.fr

Mean field games (MFG for short) are differential games, with a continuum of players, usually considered all indistinguishable and all negligible. We typically consider congestion games (i.e., agents try to avoid the regions with high concentrations), where we look for a Nash equilibrium, to be translated into a system of PDEs.

MFG theory is now a very lively topic, and the literature is rapidly growing. Among the references for a general overview of the original developments of this theory, we recommend the videotapes of the 6-year course given by P.-L. Lions at Collège de France [25] and the lecture notes by P. Cardaliaguet [11], directly inspired from these courses.

The initial goal behind the theory is to study the limit as $N \rightarrow \infty$ of N -players games, each player choosing a trajectory $x_i(t)$ and optimizing a quantity

$$\int_0^T \left(\frac{|x'_i(t)|^2}{2} + g_i(x_1(t), \dots, x_N(t)) \right) dt + \Psi_i(x_i(T)).$$

In particular, we are interested in the case where g_i penalizes points close to too many other players x_j , $j \neq i$. The indistinguishability assumptions translates into the fact that all the functions Ψ_i are equal and the cost g_i takes the form

$$g_i(x_1(t), \dots, x_N(t)) = g \left(x_i, \frac{1}{N-1} \sum_{j \neq i} \delta_{x_j} \right),$$

which means that the congestion cost felt by each agent only depends on his position compared to the distribution of the other players, i.e., the probability measure $\frac{1}{N-1} \sum_{j \neq i} \delta_{x_j}$. In the limit as $N \rightarrow \infty$, this measure is essentially the same as $\rho = \frac{1}{N} \sum_{j=1}^N \delta_{x_j}$, which gives a cost of the form $g(x, \rho)$.

Many possible dependences can be considered, but the main one that we will consider in this paper is a local congestion cost which takes the form, in the continuous limit, of $g(x, \rho(x))$, for a function $g : \Omega \times \mathbb{R}^+ \rightarrow \overline{\mathbb{R}}$, increasing in its second variable. Note that from the mathematical point of view, this is the most intriguing choice, as non-local congestion costs (of the form $g(x, (K * \rho)(x))$) for an interaction kernel K , so that the effective density perceived by the agents is of the form $\int K(x - y)\rho(y)dy$ automatically provide more compactness and regularity which are not available for local costs. We refer to [12] for rigorous definitions and results for the local case.

Rigorous convergence results starting from N players and letting $N \rightarrow \infty$ in the previous differential game are a delicate issue, beyond the scope of the present paper, the short presentation above only aimed at introducing the continuous version that we will detail in the sequel of the paper.

1.1 A Coupled System of PDEs

We will describe now in a more precise way the continuous equilibrium problem resulting from the previous considerations. We consider a population of agents moving inside Ω (which can be a domain in \mathbb{R}^d , the flat torus $\mathbb{T}^d := \mathbb{R}^d / \mathbb{Z}^d \dots$), and we suppose that every agent chooses his own trajectory solving

$$\min \int_0^T \left(\frac{|x'(t)|^2}{2} + g(x, \rho_t(x(t))) \right) dt + \Psi(x(T)),$$

with given initial point $x(0)$; here, g is a given increasing function of the density ρ_t at time t . The agent hence tries to avoid overcrowded regions.

For the moment, we consider the evolution of the density ρ_t as an input, i.e., we suppose that agents know it. Supposing the function $h(t, x) = g(\rho_t(x))$ to be given, a crucial tool to study the above individual optimization problem is the value function. The value function φ for this problem is defined as

$$\varphi(t_0, x_0) := \min \left\{ \int_{t_0}^T \left(\frac{|x'(t)|^2}{2} + h(t, x) \right) dt + \Psi(x(T)), \quad x : [t_0, T] \rightarrow \Omega, x(t_0) = x_0 \right\},$$

where, again, $h(t, x) = g(x, \rho_t(x))$.

Dynamic programming arguments from optimal control theory provides useful information on the role of the value function. First, we know that it solves a Hamilton–Jacobi equation

$$(HJ) \quad \begin{cases} -\partial_t \varphi + \frac{1}{2} |\nabla \varphi|^2 = h, \\ \varphi(T, x) = \Psi(x). \end{cases}$$

This equation is to be intended in the viscosity sense, but the presentation in this section will be quite formal. The important point for the moment is that the value function φ depends, through a Hamilton–Jacobi equation, on Ψ and h .

Moreover, the optimal trajectories $x(t)$ in the above control problem can be computed using the value function. Indeed, the optimal trajectories are the solution of

$$x'(t) = -\nabla \varphi(t, x(t))$$

(we do not discuss here whether these solutions are unique or not, as it depends on regularity issues on $\nabla \varphi$).

Now, given the initial density of the population ρ_0 , if we know that the agents move along solutions of an equation $x'(t) = v_t(x(t))$, an easy computation which is standard in fluid mechanics gives the PDE which is solved by the density as a function of (t, x) . This PDE is the so-called continuity equation:

$$(CE) \quad \partial_t \rho + \nabla \cdot (\rho v) = 0.$$

This equation has to be interpreted in a weak sense (see Equation (2)), and it is completed by no-flux boundary conditions $\rho v \cdot n = 0$, which model the fact that no mass enters or exits Ω .

In MFG, as standard in non-cooperative games, we look for a stable configuration, which is an equilibrium in the sense of Nash equilibria: A configuration where, keeping into account the choices of the others, no player would spontaneously decide to change his own choice.

This means that we can consider the densities ρ_t as an input, compute the optimal trajectories, which depend on $h = g(x, \rho_t)$ through the (HJ) equation, then compute the solution of (CE) and get new densities as an output: The configuration is an equilibrium if and only if the output densities coincide with the input. Alternatively, we can consider the trajectories of the players as an input, compute the densities using (CE), then compute the optimal trajectories as an output via (HJ): again, the configuration is an equilibrium if and only input=output.

All in all, an equilibrium is characterized by the following coupled system (HJ)+(CE): the function φ solves (HJ) with a right-hand side depending on ρ , which on turn evolves according to (CE) with a velocity field depending on $\nabla\varphi(t, x)$.

$$\begin{cases} -\partial_t \varphi + \frac{|\nabla \varphi|^2}{2} = g(x, \rho), \\ \partial_t \rho - \nabla \cdot (\rho \nabla \varphi) = 0, \\ \varphi(T, x) = \Psi(x), \quad \rho(0, x) = \rho_0(x). \end{cases} \quad (1)$$

Later (Section 5), we will see how to define a similar approach for the stochastic case, which means that agents follow controlled stochastic differential equations of the form $dX_t = \alpha_t dt + \sqrt{2}dW_t$ and minimize $\mathbb{E}[\int_0^T (\frac{1}{2}\alpha_t^2 + g(X_t, \rho_t(X_t))dt)]$. In this case, a Laplacian appears both in the (HJ) and in the (CE) equations:

$$-\partial_t \varphi - \Delta \varphi + \frac{|\nabla \varphi|^2}{2} - g(\rho) = 0, \quad \partial_t \rho - \Delta \rho - \nabla \cdot (\rho \nabla \varphi) = 0.$$

1.2 Variational Principle

It happens that a solution to the equilibrium system (1) can be found by an overall minimization problem as first outlined in the seminal work of Lasry and Lions [23]. We consider all the possible population evolutions, i.e., pairs (ρ, v) satisfying $\partial_t \rho + \nabla \cdot (\rho v) = 0$ (note that this is the Eulerian way of describing such a movement; in Section 3, we will see how to express it in a Lagrangian language) and we minimize the following energy

$$\mathcal{A}(\rho, v) := \int_0^T \int_{\Omega} \left(\frac{1}{2} \rho_t |v_t|^2 + G(x, \rho_t) \right) dx dt + \int_{\Omega} \Psi d\rho_T,$$

where G is the anti-derivative of g with respect to its second variable, i.e., $\partial_s G(x, s) = g(x, s)$ for $s \in \mathbb{R}^+$ with $G(x, 0) = 0$. We fix by convention $G(x, s) = +\infty$ for $s < 0$. Note in particular that G is convex in its second variable, as its derivative is the increasing function g .

The above minimization problem recalls the Benamou–Brenier dynamic formulation for optimal transport (see [6]). The main difference with the Benamou–Brenier problem is that here we add to the kinetic energy a congestion cost G ; also note that usually in optimal transport the target measure ρ_T is fixed, and here it is part of the optimization (but this is not a crucial difference). Finally, note that the minimization of a Benamou–Brenier energy with a congestion cost was already present in [10] where the congestion term was used to model the motion of a crowd with panic.

As is often the case in congestion games, the quantity $\mathcal{A}(\rho, v)$ is not the total cost for all the agents. Indeed, the term $\int \int \frac{1}{2} \rho |v|^2$ is exactly the total kinetic energy, and the last term $\int \Psi d\rho_T$ is the total final cost, but the term $\int G(x, \rho)$ is not the total congestion cost, which should be instead $\int \rho g(x, \rho)$. This means that the equilibrium minimizes an overall energy (we have what is called a potential game), but not the total cost; which gives rise to the so-called *price of anarchy*.

Another important point is the fact that the above minimization problem is convex, which was by the way the key idea of [6]. Indeed, the problem is not convex in the variables (ρ, v) , because of the product term $\rho |v|^2$ in the functional and of the product ρv in the differential constraint. But if one changes variable, defining $w = \rho v$ and using the variables (ρ, w) , then the constraint becomes linear and the functional convex. We will write $\bar{\mathcal{A}}(\rho, w)$ for the functional $\mathcal{A}(\rho, v)$ written in these variables. The important point for convexity is that the function

$$\mathbb{R} \times \mathbb{R}^d \ni (s, w) \mapsto \begin{cases} \frac{|w|^2}{2s} & \text{if } s > 0, \\ 0 & \text{if } (s, w) = (0, 0), \\ +\infty & \text{otherwise} \end{cases}$$

is convex (and it is actually obtained as $\sup\{as + b \cdot w : a + \frac{1}{2}|b|^2 \leq 0\}$).

This will be the basis for the numerical method of Section 6.

In order to convince the reader of the connection between the minimization of $\mathcal{A}(\rho, v)$ (or of $\bar{\mathcal{A}}(\rho, w)$) and the equilibrium system (1), we will use some formal argument from convex duality. We will see in Section 2 how to rigorously justify this equivalence.

In order to formally produce a dual problem to $\min \mathcal{A}$, we will use a min–max exchange procedure. First, we write the constraint $\partial_t \rho + \nabla \cdot (\rho v) = 0$ in weak form, i.e.,

$$\int_0^T \int_{\Omega} (\rho \partial_t \phi + \nabla \phi \cdot \rho v) + \int_{\Omega} \phi_0 \rho_0 - \int_{\Omega} \phi_T \rho_T = 0 \quad (2)$$

for every function $\phi \in C^1([0, T] \times \Omega)$ (note that we do not impose conditions on the values of ϕ on $\partial\Omega$, hence this is equivalent to completing (CE) with a no-flux boundary condition $\rho v \cdot n = 0$). Also note that, if ρ_0 is a datum of our problem, ρ_T

is not, and the Equation (2) does not make sense unless we give a meaning at ρ_t for every instant of time t . This will be done in Section 3 (see Proposition 3.1), where we will interpret ρ_t as a(n absolutely) continuous curve in the space of measures. However, we do not insist on this now, as the presentation stays quite formal.

Using (2), we can rewrite our problem as

$$\min_{\rho, v} \mathcal{A}(\rho, v) + \sup_{\phi} \int_0^T \int_{\Omega} (\rho \partial_t \phi + \nabla \phi \cdot \rho v) + \int_{\Omega} \phi_0 \rho_0 - \int_{\Omega} \phi_T \rho_T,$$

since the sup in ϕ takes value 0 if the constraint is satisfied and $+\infty$ if not. We now switch the inf and the sup and get

$$\sup_{\phi} \int_{\Omega} \phi_0 \rho_0 + \inf_{\rho, v} \int_{\Omega} (\Psi - \phi_T) \rho_T + \int_0^T \int_{\Omega} \left(\frac{1}{2} \rho_t |v_t|^2 + G(x, \rho_t) + \rho \partial_t \phi + \nabla \phi \cdot \rho v \right) dx dt.$$

First, we minimize w.r.t. v , thus obtaining $v = -\nabla \phi$ and we replace $\frac{1}{2} \rho |v|^2 + \nabla \phi \cdot \rho v$ with $-\frac{1}{2} \rho |\nabla \phi|^2$. Then we get, in the double integral,

$$\inf_{\rho} \{G(x, \rho) - \rho(-\partial_t \phi + \frac{1}{2} |\nabla \phi|^2)\} = -\sup_{\rho} \{p\rho - G(x, \rho)\} = -G^*(x, p),$$

where we set $p := -\partial_t \phi + \frac{1}{2} |\nabla \phi|^2$ and G^* is defined as a Legendre transform with respect to its second variable only. Then, we observe that the minimization in the final cost simply gives as a result 0 if $\Psi \geq \phi_T$ (since the minimization is only performed among positive ρ_T) and $-\infty$ otherwise. Hence, we obtain a dual problem of the form

$$\sup \left\{ -\mathcal{B}(\phi, p) := \int_{\Omega} \phi_0 \rho_0 - \int_0^T \int_{\Omega} G^*(x, p) : \phi_T \leq \Psi, -\partial_t \phi + \frac{1}{2} |\nabla \phi|^2 = p \right\}.$$

Note that the condition $G(x, \rho) = +\infty$ for $\rho < 0$ implies $G^*(x, p) = 0$ for $p \leq 0$. This in particular means that in the above maximization problem, one can suppose $p \geq 0$ (indeed, replacing p with p_+ does not change the G^* part, but improves the value of ϕ_0 , considered as a function depending on p). The choice of using two variables (ϕ, p) connected by a PDE constraint instead of only ϕ is purely conventional, and it allows for a dual problem which has a particular symmetry w.r.t. the primal one. Also the choice of the sign is conventional and due to the computation that we will perform later (in particular in Section 4).

Now, standard arguments in convex duality, which will be made precise in the next section, allow us to say that optimal pairs (ρ, v) are obtained by looking at saddle points $((\rho, v), (\phi, p))$. This means that, whenever (ρ, v) minimizes \mathcal{A} , then there exists a pair (ϕ, p) , solution of the dual problem, such that

- v minimizes $\frac{1}{2} \rho |v|^2 + \nabla \phi \cdot \rho v$, i.e., $v = -\nabla \phi$ ρ -a.e. This gives (CE): $\partial_t \rho - \nabla \cdot (\rho \nabla \phi) = 0$.
- ρ minimizes $G(x, \rho) - p\rho$, i.e., $g(x, \rho) = p$ if $\rho > 0$ or $g(x, \rho) \geq p$ if $\rho = 0$ (so that $g(x, \rho) = p_+$); this gives (HJ): $-\partial_t \phi + \frac{1}{2} |\nabla \phi|^2 = g(x, \rho)$ on $\{\rho > 0\}$ (as the

reader can see, there are some subtleties where the mass ρ vanishes; this will be discussed later).

- ρ_T minimizes $(\Psi - \phi_T)\rho_T$ among $\rho_T \geq 0$. But this is not a condition on ρ_T , but rather on ϕ_T : we must have $\phi_T = \Psi$ on $\{\rho_T > 0\}$, otherwise there is no minimizer. This gives the final condition in (HJ).

This provides an informal justification for the equivalence between the equilibrium and the global optimization. What we lack for the moment is the fact that there is no duality gap between $\min \mathcal{A}$ and $\max -\mathcal{B}$, and that there is existence of minimizers (in particular in the dual problem, and in which spaces). Also, even once these issues are clarified, what we will get will only be a very weak solution to the coupled system (CE)+(HJ). Nothing guarantees that this solution actually encodes the individual minimization problem of each agent. This will be clarified in Section 3 where a Lagrangian point of view will be presented.

We finish this section with a last variant, inspired by the crowd motion model of [26]. We would like to consider a variant where, instead of adding a penalization $g(x, \rho)$, we impose a capacity constraint $\rho \leq 1$. How to give a proper definition of equilibrium? A first, naive, idea would be the following: when $(\rho_t)_t$ is given, every agent minimizes his own cost paying attention to the constraint $\rho_t(x(t)) \leq 1$. But if ρ already satisfies $\rho \leq 1$, then the choice of only one extra agent will not violate the constraint (since we have a non-atomic game), and the constraint becomes empty. As already pointed out in [29], this cannot be the correct definition.

In [29] an alternative model is formally proposed, but no solution has been given to it so far, and it is likely to be non-variational. Instead, we can look at the variational problem

$$\min \left\{ \int_0^T \int_{\Omega} \frac{1}{2} \rho_t |v_t|^2 + \int_{\Omega} \Psi \rho_T : \rho \leq 1 \right\}.$$

This means using $G(\rho) = 0$ for $\rho \in [0, 1]$ and $+\infty$ otherwise. The dual problem can be computed and we obtain

$$\sup \left\{ \int_{\Omega} \phi_0 \rho_0 - \int_0^T \int_{\Omega} p_+ : \phi_T \leq \Psi, p \geq 0, -\partial_t \phi + \frac{1}{2} |\nabla \phi|^2 = p \right\}$$

(note that this problem is also obtained as the limit $m \rightarrow \infty$ of $g(\rho) = \rho^m$; indeed the functional $\frac{1}{m+1} \int \rho^{m+1}$ Γ -converges to the constraint $\rho \leq 1$ as $m \rightarrow \infty$).

By looking at the primal-dual optimality conditions, we get again $v = -\nabla \phi$ and $\phi_T = \Psi$, but the optimality of ρ means

$$0 \leq \rho < 1 \Rightarrow p = 0, \quad \rho = 1 \Rightarrow p \geq 0.$$

This gives the following MFG system

$$\begin{cases} -\partial_t \varphi + \frac{|\nabla \varphi|^2}{2} = p, \\ \partial_t \rho - \nabla \cdot (\rho \nabla \varphi) = 0, \\ p \geq 0, \rho \leq 1, p(1-\rho) = 0, \\ \varphi(T, x) = \Psi(x), \quad \rho(0, x) = \rho_0(x). \end{cases}$$

Formally, by looking back at the relation between (HJ) and optimal trajectories, we can guess that each agent solves

$$\min \int_0^T \left(\frac{|x'(t)|^2}{2} + p(t, x(t)) \right) dt + \Psi(x(T)). \quad (3)$$

Here, p is a pressure arising from the incompressibility constraint $\rho \leq 1$ and only present in the saturated zone $\{\rho = 1\}$, and it finally acts as a price paid by the agents to travel through saturated regions. From the economical point of view, this is meaningful: due to a capacity constraint, the most attractive regions develop a positive price to be paid to pass through them, and this price is such that, if the agents keep it into account in their choices, then their mass distribution will indeed satisfy the capacity constraints.

This problem has been studied in [16], where suitable regularity results (see also Section 4) allow one to give a meaning to what we said above. This is necessary, because of the fact that, from the linear growth in the dual problem, we should not a priori expect p to be better than a measure, and this makes it difficult to define the integral over the trajectory in (3). The only way to handle this difficulty is to prove extra summability for p .

2 Existence of Minimizers and Convex Duality

The aim of this section is to explain how to relate rigorously the two variational problems formally introduced in Section 1.2 with a suitable weak notion of solutions for the MFG system, following the approach developed by Cardaliaguet [12] (also see [13]) and further refined in Cardaliaguet and Graber [14]. In what follows, the spatial domain Ω will denote either a smooth bounded domain of \mathbb{R}^d or the flat torus $\Omega := \mathbb{T}^d = \mathbb{R}^d \setminus \mathbb{Z}^d$ (periodic case). To fix ideas, we take a quadratic Hamiltonian, as in Section 1 and assume that the congestion term G , given by $G(x, \rho) := \int_0^\rho g(x, s) ds$ for $\rho \geq 0$ and $G(x, \rho) = +\infty$ if $\rho < 0$, is superlinear

$$\frac{G(x, \rho)}{\rho} \rightarrow +\infty \text{ as } \rho \rightarrow \infty \quad (4)$$

uniformly in x (which in particular includes the case of a *hard* congestion constraint, i.e., $G(x, \rho) = 0$ if $\rho \in [0, 1]$ and $+\infty$ otherwise). Due to the dependence on x (which is non-essential in most of the paper, but some problems could become trivial

without, in particular in Section 5), we also add this very mild assumption

$$x \mapsto G(x, \rho) \text{ is l.s.c. for all } \rho. \quad (5)$$

Recall that the initial distribution of players is given and is denoted $\rho_0 \in \mathcal{P}(\Omega)$ as well as the terminal cost $\Psi \in C(\Omega)$. Let us then consider the variational problem

$$\inf_{\phi \in C^1([0, T] \times \Omega), \phi(T, \cdot) \leq \Psi} \int_0^T \int_{\Omega} G^* \left(x, -\partial_t \phi + \frac{1}{2} |\nabla \phi|^2 \right) dx dt - \int_{\Omega} \phi(0, x) d\rho_0(x). \quad (6)$$

Since $G^*(x, p)$ is non-decreasing in p and identically 0 for $p \in (-\infty, 0)$, this is a convex minimization problem, which can be rewritten (as we did in Section 1) as

$$\inf_{(\phi, p) \in \mathcal{F}} \int_0^T \int_{\Omega} G^*(x, p) dx dt - \int_{\Omega} \phi(0, x) d\rho_0(x),$$

where \mathcal{F} consists of all pairs $(\phi, p) \in C^1([0, T] \times \Omega) \times C^0([0, T] \times \Omega)$ such that $\phi(T, \cdot) \leq \Psi$ and

$$-\partial_t \phi + \frac{1}{2} |\nabla \phi|^2 \leq p \quad (7)$$

(note, that imposing equality or inequality in (13) is the same in the above minimization problem, since G^* is non-decreasing).

The dual of this problem is (see [19]) is

$$\inf_{(\rho, w) \in \mathcal{K}} \bar{\mathcal{A}}(\rho, w) := \left\{ \int_0^T \int_{\Omega} \left(\frac{1}{2} \frac{|w_t(x)|^2}{\rho_t(x)} + G(x, \rho_t(x)) \right) dx dt + \int_{\Omega} \Psi \rho_T \right\} \quad (8)$$

where \mathcal{K} consists of all pairs $(t, x) \mapsto (\rho_t(x), w_t(x))$ in $L^1((0, T) \times \Omega) \times \mathcal{M}((0, T) \times \Omega, \mathbb{R}^d)$ such that each ρ_t is a probability measure and

$$\partial_t \rho + \nabla \cdot w = 0, \quad \rho(0, \cdot) = \rho_0 \quad (9)$$

in the sense of distributions (with respect to the formulation given in terms of the continuity equation (CE) in Section 1, the formulation above corresponds to the change of variable $(\rho, v) \mapsto (\rho, w) = (\rho, \rho v)$ that we already mentioned in the spirit of the Benamou–Brenier formulation of optimal transport and which makes the minimization problem (8) convex). More precisely, the Fenchel–Rockafellar duality theorem (see [19]) gives:

Theorem 2.1 Suppose (4) and (5), then we have

$$\min(8) = -\inf(6) \quad (10)$$

The proof of this theorem can be obtained following the same arguments as in [12].

In particular, a minimizer to (8) exists. One further has uniqueness of such a minimizer (ρ, w) if G is strictly convex. Since one cannot expect that a smooth minimizer to (6) exists, one has to suitably relax (6). To do so, following [14], we shall further assume that for some exponent $q > 1$, and some constant $C > 0$, one has:

$$\frac{1}{C}\rho^q - C \leq G(x, \rho) \leq C\rho^q + C \quad (11)$$

for every $\rho \geq 0$ so that G^* satisfies a similar power growth condition with the dual exponent $q' = q/(q-1)$. The relaxation of (6) is then as follows:

$$\inf_{(\phi, p) \in \tilde{\mathcal{F}}} \mathcal{B}(\phi, p) := \int_0^T \int_{\Omega} G^*(x, p(t, x)) dx dt - \int_{\Omega} \phi(0, x) d\rho_0(x) \quad (12)$$

where $\tilde{\mathcal{F}}$ consists of all pairs $(\phi, p) \in \text{BV}((0, T) \times \Omega) \times L^1((0, T) \times \Omega)$ such that $\nabla\phi \in L^2((0, T) \times \Omega)$, $p_+ \in L^q((0, T) \times \Omega)$, $\phi(T, \cdot) \leq \Psi$ in the sense of traces and

$$-\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 \leq p \quad (13)$$

in the sense of distributions. As shown in [14], Problem (12) is really a relaxation of (6) in the sense that the values of both problems coincide. The existence of a minimizer to the relaxed problem (12) is, however, more involved and requires more assumptions, a key point is to understand how an $L^{q'}$ (with $q' = q/(q-1)$) bound on p gives pointwise bounds on ϕ subsolution of the HJ equation (13). Such bounds can be obtained (see Lemma 2.7 in [14]) and subsequently the existence of a minimizer for (12) can be proved as soon as

$$\rho_0 \in C(\overline{\Omega}), \min_{\overline{\Omega}} \rho_0 > 0, \Psi \in W^{1,\infty}(\Omega), q < 1 + \frac{2}{d}. \quad (14)$$

We mention also the work in [16], corresponding to the constrained case $\rho \leq 1$ (hence, in some sense, to $q = \infty$), where a similar result is proven under the assumption $\|\rho_0\|_{L^\infty} < 1$.

To sum up, a rigorous existence result established by Cardaliaguet and Graber [14] (also see [13] for a slightly different but related problem) can be summarized as:

Theorem 2.2 *Assume (11)–(14), then the infimum in (12) is achieved and coincides with $-\min(8)$.*

This said, it remains to understand in which sense the duality relation $0 = \mathcal{B}(\phi, p) + \tilde{\mathcal{A}}(\rho, w)$ relating an optimal $(\rho, w) \in \mathcal{K}$ for (8) to an optimal $(\phi, p) \in \tilde{\mathcal{F}}$ for the relaxed problem gives rise to a mean-field-game-like system. This is the object of the main result for which we refer again to [12] and [14]:

Theorem 2.3 Assume (11)–(14) and let $(\rho, w) \in \mathcal{K}$ solve (8) and $(\phi, p) \in \tilde{\mathcal{F}}$ solve (12), then

$$\partial_t \rho - \nabla \cdot (\rho \nabla \phi) = 0 \text{ in } \mathcal{D}'((0, T) \times \Omega), \quad \rho(0, \cdot) = \rho_0, \quad (15)$$

$$-\partial_t \phi + \frac{1}{2} |\nabla \phi|^2 \leq g(x, \rho) \text{ in } \mathcal{D}'((0, T) \times \Omega), \quad \phi(T, \cdot) \leq \Psi, \quad (16)$$

$$\begin{aligned} & \int_0^T \int_{\Omega} \rho_t(x) \left(\frac{1}{2} |\nabla \phi(t, x)|^2 - g(x, \rho_t(x)) \right) dx dt \\ &= \int_{\Omega} \Psi(x) \rho(T, x) dx - \int_{\Omega} \phi(0, x) \rho_0(x) dx. \end{aligned} \quad (17)$$

It is worth pointing out that (15)–(16)–(17) imply that the Hamilton–Jacobi equation is satisfied in the following weak sense: ρ -a.e one has

$$-(\partial_t \phi)^{\text{ac}} + \frac{1}{2} |\nabla \phi|^2 = g(x, \rho)$$

where $(\partial_t \phi)^{\text{ac}}$ denotes the absolutely continuous part of the measure $\partial_t \phi$. In other words, if ρ vanishes nowhere and $\partial_t \phi$ has no singular part then (ρ, ϕ) solves the MFG system in some appropriate weak sense and the MFG system actually is a necessary and sufficient optimality condition for the variational problems in duality (8)–(12).

3 The Lagrangian Framework

In this section, we present an alternative point of view for the overall minimization problem presented in the previous sections. As far as now, we only looked at an Eulerian point of view, where the motion of the population is described by means of its density ρ and of its velocity field v . The Lagrangian point of view would be, instead, to describe the motion by describing the trajectory of each agent. Since the agents are supposed to be indistinguishable, then we just need to determine, for each possible trajectory, the number of agents following it (and not their names...); this means looking at a measure on the set of possible paths.

Set $\mathcal{C} = H^1([0, T]; \Omega)$; this will be the space of possible paths that we use. In general, absolutely continuous paths would be the good choice, but we can restrict our attention to H^1 paths because of the kinetic energy term that we have in our minimization. We define the evaluation maps $e_t : \mathcal{C} \rightarrow \Omega$, given for every $t \in [0, T]$ by $e_t(\omega) = \omega(t)$. Also, we define the kinetic energy functional $K : \mathcal{C} \rightarrow \mathbb{R}$ given by

$$K(\omega) = \frac{1}{2} \int_0^T |\omega'|^2(t) dt.$$

We endow the space \mathcal{C} with the uniform convergence (and not the strong H^1 convergence, so that we have compactness of the sublevel sets of K).

To pass from the Eulerian to the Lagrangian framework, we will need some easy tools from optimal transport. First, we give some definitions. We refer to [30] (Chapters 1, 5, and 7) and to [5, 32] for more details and complete proofs.

Given two probability measures $\mu, \nu \in \mathcal{P}(\Omega)$, we consider the set of transport plans

$$\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_x)_\# \gamma = \mu, (\pi_y)_\# \gamma = \nu, \}$$

i.e., those probability measures on the product space having μ and ν as marginal measures.

We consider the minimization problem

$$\min \left\{ \int |x - y|^2 d\gamma : \gamma \in \Pi(\mu, \nu) \right\},$$

which is called the Kantorovich optimal transport problem for the cost $c(x, y) = |x - y|^2$ from μ to ν .

The value of this minimization problem with the quadratic cost may also be used to define a quantity called Wasserstein distance:

$$W_2(\mu, \nu) := \sqrt{\min \left\{ \int |x - y|^2 d\gamma : \gamma \in \Pi(\mu, \nu) \right\}}.$$

If we suppose that Ω is compact, this quantity may be proven to be a distance over $\mathcal{P}(\Omega)$, and it metrizes the weak-* convergence of probability measures. The space $\mathcal{P}(\Omega)$ endowed with the distance W_2 is called Wasserstein space of order 2 and denoted in this paper by $\mathbb{W}_2(\Omega)$.

We summarize here below how the theory of optimal transport helps in studying the relation between curves of measures and measures of curves, which is the main point in passing from the Eulerian to the Lagrangian formalism.

We recall the definition of metric derivative in metric spaces, applied to the case of $\mathbb{W}_2(\Omega)$: for a curve $t \mapsto \rho_t \in \mathbb{W}_2(\Omega)$, we define

$$|\rho'| (t) := \lim_{s \rightarrow t} \frac{W_2(\rho_s, \rho_t)}{|s - t|},$$

whenever this limit exists. If the curve $t \mapsto \rho_t$ is absolutely continuous for the W_2 distance, then this limit exists for a.e. t . The important fact, coming from the Benamou–Brenier formula and explained for the first time in [5] (see also Chapter 5 in [30]) is that the absolutely continuous curves in $\mathbb{W}_2(\Omega)$ are exactly those curves which admit the existence of a velocity field v_t solving (CE) together with ρ and that the metric derivative $|\rho'| (t)$ can be computed as the minimal norm $\|v_t\|_{L^2(\rho_t)}$ among those vector fields. This is part of the following statement.

Proposition 3.1 Suppose (ρ, v) satisfies the continuity equation $\partial_t \rho + \nabla \cdot (\rho v) = 0$ and $\int_0^T \int_{\Omega} \rho |v|^2 < \infty$. Then, there exists a representative of ρ such that $t \mapsto \rho_t \in \mathbb{W}_2(\Omega)$ is absolutely continuous and $|\rho'|_t(t) \leq ||v_t||_{L^2(\rho_t)} a.e.$ Moreover, there exists a probability measure $Q \in \mathcal{P}(\mathcal{C})$ such that $\rho_t = (e_t)_\# Q$ and

$$\int_{\mathcal{C}} K(\omega) dQ(\omega) \leq \frac{1}{2} \int_0^T \int_{\Omega} \rho |v|^2.$$

Conversely, if $\rho_t = (e_t)_\# Q$ for a probability measure $Q \in \mathcal{P}(\mathcal{C})$ with $\int_{\mathcal{C}} K(\omega) dQ(\omega) < \infty$, then $t \mapsto \rho_t \in \mathbb{W}_2(\Omega)$ is absolutely continuous and there exists a time-dependent family of vector fields $v_t \in L^2(\rho_t)$ such that $\partial_t \rho + \nabla \cdot (\rho v) = 0$ and

$$\frac{1}{2} \int_0^T \int_{\Omega} \rho |v|^2 \leq \int_{\mathcal{C}} K(\omega) dQ(\omega).$$

The above proposition (whose proof can be found combining, for instance, Theorems 5.14 and 5.31 in [30]) allows to rewrite the minimization problem

$$\min \{\mathcal{A}(\rho, v) : \partial_t \rho + \nabla \cdot (\rho v) = 0\},$$

in the following form:

$$\min \left\{ J(Q) := \int_{\mathcal{C}} K dQ + \int_0^T \mathcal{G}((e_t)_\# Q) dt + \int_{\Omega} \Psi d(e_T)_\# Q, \quad Q \in \mathcal{P}(\mathcal{C}), (e_0)_\# Q = \rho_0 \right\}, \quad (18)$$

where $\mathcal{G} : \mathcal{P}(\Omega) \rightarrow \overline{\mathbb{R}}$ is defined through

$$\mathcal{G}(\rho) = \begin{cases} \int G(x, \rho(x)) dx & \text{if } \rho \ll \mathcal{L}^d, \\ +\infty & \text{otherwise.} \end{cases}$$

The functional \mathcal{G} is a typical local functional defined on measures (see [8]). It is lower-semicontinuous w.r.t. weak convergence of probability measures provided $\lim_{\rho \rightarrow \infty} G(x, \rho)/\rho = +\infty$ (which is the same as $\lim_{\rho \rightarrow \infty} g(x, \rho) = +\infty$), see, for instance, Proposition 7.7 in [30].

Under these assumptions, it is easy to prove, by standard semicontinuity arguments in the space $\mathcal{P}(\mathcal{C})$, that a minimizer of (18) exists. We summarize this fact, together with the corresponding optimality conditions, in the next proposition.

Proposition 3.2 Suppose that Ω is compact and that G is a convex function satisfying (4) and (5). Then, the problem (18) admits a solution \bar{Q} .

Moreover, \bar{Q} is a solution if and only if for any other competitor $Q \in \mathcal{P}(\mathcal{C})$ with $J(Q) < +\infty$ with $(e_0)_\# Q = \rho_0$ we have

$$J_h(Q) \geq J_h(\bar{Q}),$$

where J_h is the linear functional

$$J_h(Q) = \int K dQ + \int_0^T \int_{\Omega} h(t, x) d(e_t)_\# Q + \int_{\Omega} \Psi d(e_T)_\# Q,$$

the function h being defined through $\rho_t = (e_t)_\# \bar{Q}$ and $h(t, x) = g(x, \rho_t(x))$.

Remark 1 The above optimality condition and the interpretation in terms of equilibria that we will give below are very close to what has been studied in the framework of *continuous Wardrop equilibria* in [17] (see also [18] for a survey of the theory). Indeed, in such a framework, we associate with each measure Q on \mathcal{C} a traffic intensity i_Q (which is a measure on Ω), and we define a weighted length on curves ω using i_Q as a weighting factor. We then prove that the measure Q which minimizes a suitable functional (also constructed via the anti-derivative of a congestion function g) minimizes its linearization, which in turn implies that the same Q is concentrated on curves which are geodesic for this weighted length, which depends on Q itself! The analogy is very strict, with the only difference that the framework of Wardrop equilibria (which are traditionally studied in a discrete framework on networks, see [33]) are a statical object. The use of time to parametrize curves in Wardrop models is fictitious, and one has to think at a continuous traffic flows, where mass is constantly injected in some parts of Ω and absorbed in other parts (see Chapter 4 of [30] for a general picture of this kind of models).

We now consider the functional J_h . Note that the function h is obtained from the densities ρ_t , which means that it is well-defined a.e. But the integral $\int_0^T \int_{\Omega} h(t, x) d(e_t)_\# Q$ is well defined and does not depend on the representative of h , since $J(Q) < +\infty$ implies that all the measures $(e_t)_\# Q$ are absolutely continuous. Hence, this functional is well defined for $h \geq 0$ measurable.

Yet, if we suppose for a while that h is a continuous function, we can also write

$$\int_0^T \int_{\Omega} h(t, x) d(e_t)_\# Q = \int_{\mathcal{C}} dQ \int_0^T h(t, \omega(t)) dt$$

and hence we get

$$J_h(Q) = \int_{\mathcal{C}} dQ(\omega) \left(K(\omega) + \int_0^T h(t, \omega(t)) dt + \Psi(\omega(T)) \right).$$

It is not difficult to see that in this case \bar{Q} satisfies the optimality conditions of Proposition 3.2 if and only if Q -a.e. curve ω satisfies

$$\begin{aligned} K(\omega) + \int_0^T h(t, \omega(t)) dt + \Psi(\omega(T)) &\leq K(\tilde{\omega}) \\ &+ \int_0^T h(t, \tilde{\omega}(t)) dt + \Psi(\tilde{\omega}(T)) \quad \text{for every } \tilde{\omega} \text{ s.t. } \tilde{\omega}(0) = \omega(0). \end{aligned}$$

This is exactly the equilibrium condition in the MFG! Indeed, the MFG equilibrium condition can be expressed in Lagrangian language in the following way: find Q such that, if we define $\rho_t = (e_t)_\# \bar{Q}$ and $h(t, x) = g(\rho_t(x))$, then Q is concentrated on minimizers of

$$A_h(\omega, [0, T]) := K(\omega) + \int_0^T h(t, \omega(t)) dt + \Psi(\omega(T))$$

for fixed initial point let us also define

$$\begin{aligned} A_h(\omega, [t_0, T]) &:= \int_{t_0}^T \frac{1}{2} |\omega'|^2(t) dt + \int_{t_0}^T h(t, \omega(t)) dt + \Psi(\omega(T)), \\ A_h(\omega, [t_0, t_1]) &:= \int_{t_0}^{t_1} \frac{1}{2} |\omega'|^2(t) dt + \int_{t_0}^{t_1} h(t, \omega(t)) dt \quad \text{for } t_1 < T. \end{aligned}$$

Here, we just found out that \bar{Q} satisfies this equilibrium condition if and only if it minimizes J .

The question which thus arises is how to give a rigorous meaning to this equilibrium condition when $h \notin C^0$. We will not enter details here, but we want to stress that there is a solution which passes through the choice of a precise representative of h . Indeed, following what Ambrosio and Figalli did in [4] we can define $h_r(t, x) = \int_{B(x,r)} h(t, y) dy$ and $\hat{h}(x) := \limsup_{r \rightarrow 0} h_r(x)$. The technique developed in [4] (and later used in [16] for MFG with density constraints) allows to prove that if \bar{Q} minimizes J_h , then it is concentrated on curves minimizing $A_{\hat{h}}(\omega)$ by using h_r and passing to the limit as $r \rightarrow 0$, provided some upper bounds on h_r are satisfied. More precisely, if one defines Mh the maximal function of h , given by

$$Mh(t, x) = \sup_r h_r(x),$$

it is possible to prove the following.

Proposition 3.3 *Given a positive and measurable function h , suppose that \bar{Q} minimizes J_h . Then \bar{Q} is concentrated on curves ω such that, for all t_0, t_1 with $0 \leq t_0 < t_1 \leq T$,*

$$A_{\hat{h}}(\omega, [t_0, t_1]) \leq A_{\hat{h}}(\tilde{\omega}, [t_0, t_1]) \quad \text{for every } \tilde{\omega} \text{ s.t. } \tilde{\omega}(t_0) = \omega(t_0) \text{ and } \int_{t_0}^{t_1} Mh(t, \tilde{\omega}(t)) dt < +\infty.$$

In particular, this applies for $h(t, x) = g(\rho_t(x))$ whenever \bar{Q} is a solution of (18).

This condition is useful only if there are many curves $\tilde{\omega}$ satisfying $\int_{t_0}^{t_1} Mh(t, \tilde{\omega}(t)) dt < +\infty$ (note that the use of t_0 and t_1 is due to the fact that there could be only few curves such that Mh is integrable on $[0, T]$ but more such that Mh is integrable on $[t_0, t_1]$). What one can do is to take an arbitrary Q such that $J(Q) < +\infty$ and compute

$$\int \int_{t_0}^T Mh(t, \omega(t)) dt dQ(\omega) = \int_{t_0}^T \int_{\Omega} Mh(t, x) d(e_t)_\# Q dt.$$

We would like to guarantee that every Q with $J(Q) < +\infty$ is such that $\int \int_{t_0}^T Mh(t, \omega(t)) dt dQ(\omega) < \infty$. Since we know that $G(x, (e_t)_\# Q)$ is integrable, it is enough to guarantee $G^*(x, Mh) \in L^1$. In the case where $g(x, s) = s^{q-1}$, we need $Mh \in L^{q'}$. Since in this case we know $\rho \in L^q$, then $h = g(x, \rho) \in L^{q'}$ and this implies $Mh \in L^{q'}$ from standard theorems in harmonic analysis, as soon as $q' > 1$.

However, the analysis of this equilibrium condition motivates a deeper study of regularity issues, for several reasons. First, in the cases where L^∞ constraints are considered (as it happened for incompressible fluid mechanics in [4] but also in the density-constrained model of [16]), we find $q' = 1$ and we cannot get the integrability of Mh unless we first prove some extra summability. Other non-power cases (such as $g(\rho) = \log \rho$, or other) also prevent to use the L^q theory on the maximal function and require extra regularity or at least extra summability. Then, it cannot be denied that getting $h \in L^\infty$ (which implies $Mh \in L^\infty$), or even $h \in C^0$ would be much more convenient, and would allow to avoid the condition $\int_0^T Mh(t, \omega(t)) dt < \infty$ or even to use this special representative. More generally, better regularity on ρ (or on the dual variable ϕ) could give “better” solutions to the (HJ) equation (instead of just a.e. solutions).

This is why in the next section we will see a technique to prove some mild regularity results on the optimal density ρ (which are far from being complete).

4 A Bit of Regularity

We present here a technique to prove Sobolev regularity results on the solutions of (8). This technique, based on duality, is inspired from the work of [9] and has been applied to MFG in [16]. It is actually very general and [31] shows how it can be used to prove (or reprove) many regularity results in elliptic equations coming from convex variational problems.

We start from a lemma related to the duality results of Section 2.

Lemma 4.1 *For any $(\phi, p) \in \mathcal{F}$ and $(\rho, \rho v) \in \mathcal{K}$ we have*

$$\begin{aligned} \mathcal{B}(\phi, p) + \mathcal{A}(\rho, v) &= \int_{\Omega} (\Psi - \phi_T) d\rho_T + \int_0^T \int_{\Omega} (G(x, \rho) + G^*(x, p) - \rho p) dx dt \\ &\quad + \frac{1}{2} \int_0^T \int_{\Omega} \rho |v + \nabla \phi|^2 dx dt. \end{aligned}$$

Proof We start from

$$\begin{aligned} \mathcal{B}(\phi, p) + \mathcal{A}(\rho, v) &= \int_0^T \int_{\Omega} \left(\frac{1}{2} \rho |v|^2 + G(x, \rho) + G^*(x, p) \right) dx dt \\ &\quad + \int_{\Omega} \Psi d\rho_T - \int_{\Omega} \phi_0 d\rho_0. \end{aligned} \quad (19)$$

Then, we use

$$\int_{\Omega} \Psi d\rho_T - \int_{\Omega} \phi_0 d\rho_0 = \int_{\Omega} (\Psi - \phi_T) d\rho_T + \int_{\Omega} \phi_T d\rho_T - \int_{\Omega} \phi_0 d\rho_0$$

and

$$\begin{aligned} \int_{\Omega} \phi_T d\rho_T - \int_{\Omega} \phi_0 d\rho_0 &= \int_0^T \int_{\Omega} (-\phi \nabla \cdot (\rho v) + \rho \partial_t \phi) dx dt \\ &= \int_0^T \int_{\Omega} \left(\nabla \phi \cdot (\rho v) + \rho \left(\frac{1}{2} |\nabla \phi|^2 - p \right) \right) dx dt. \end{aligned}$$

If we insert this into (19) we get the desired result. \square

It is important to stress that we used the fact that ϕ is C^1 since (ρ, v) only satisfies (CE) in a weak sense, i.e., tested against C^1 functions. The same computations above would not be possible for $(\phi, p) \in \tilde{\mathcal{F}}$.

The regularity proof will come from the previous computations applied to suitable translations in space and/or time.

In order to simplify the exposition, we will choose a *spatially homogeneous* setting. In particular, we will suppose that $\Omega = \mathbb{T}^d$ is the d -dimensional flat torus, which avoids boundary issues. Also, we will suppose that g , G , and G^* do not explicitly depend on the variable x (but will explain how to adapt to the space-dependent case). To handle the case of a domain Ω with boundary, we refer to the computations in [31] which suggest how to adapt the method below. Finally, for simplicity, we will only prove in this paper local results in $(0, T)$, so that also the time boundary does not create difficulties.

Here is the intuition behind the proof in this spatially homogeneous case. First, we use Lemma 4.1 to deduce

$$\mathcal{B}(\phi, p) + \mathcal{A}(\rho, v) \geq \int_0^T \int_{\Omega} (G(x, \rho) + G^*(x, p) - \rho p) dx dt$$

(since the other terms appearing in Lemma 4.1 are positive). Then, let us suppose that there exist two functions $J, J_* : \mathbb{R} \rightarrow \mathbb{R}$ and a positive constant $c_0 > 0$ such that for all $a, b \in \mathbb{R}$ we have

$$G(a) + G^*(b) \geq ab + c_0 |J(a) - J_*(b)|^2. \quad (20)$$

Remark 2 Of course, this is always satisfied by taking $J = J_* = 0$, but there are less trivial cases. For instance, if $G(\rho) = \frac{1}{q}\rho^q$ for $q > 1$, then $G^*(p) = \frac{1}{q'}q^{r'}$, with $q' = q/(q - 1)$ and

$$\frac{1}{q}|a|^q + \frac{1}{q'}|b|^{q'} \geq ab + \frac{1}{2\max\{q, q'\}}|a^{q/2} - b^{q'/2}|^2,$$

i.e., we can use $J(a) = a^{q/2}$ and $J_*(b) = b^{q'/2}$.

We wish to show that if (ρ, v) is a minimizer of \mathcal{A} then $J(\rho) \in H_{\text{loc}}^1((0, T) \times \Omega)$. The idea is that should \mathcal{B} admit a C^1 minimizer ϕ (more precisely, a pair (ϕ, p)), then by the Duality Theorem 2.1, we have $\mathcal{B}(\phi, p) + \mathcal{A}(\rho, v) = 0$. From our assumption and Lemma 4.1, we get $J(\rho) = J_*(p)$. If we manage to show that $\tilde{\rho}(t, x) := \rho(t + \eta, x + \delta)$ with a corresponding velocity field \tilde{v} is close to minimizing \mathcal{A} , and more precisely

$$\mathcal{A}(\tilde{\rho}, \tilde{v}) \leq \mathcal{A}(\rho, v) + C(|\eta|^2 + |\delta|^2) \quad (21)$$

for small $\eta \in \mathbb{R}$, $\delta \in \mathbb{R}^d$, then we would have

$$C(|\eta|^2 + |\delta|^2) \geq \mathcal{A}(\tilde{\rho}, \tilde{v}) + \mathcal{B}(\phi, p) \geq c||J(\tilde{\rho}) - J_*(p)||_{L^2}^2.$$

However, we already know that $J_*(p) = J(\rho)$, and so

$$C(|\eta|^2 + |\delta|^2) \geq c||J(\tilde{\rho}) - J(\rho)||_{L^2}^2,$$

which would mean that $J(\rho)$ is H^1 as we have estimated the squared L^2 norm of the difference of $J(\rho)$ and its translation by the squared length of the translation. Of course, there are some technical issues that need to be taken care of, for instance $\tilde{\rho}$ is not even well defined (as we could need the value of ρ outside $[0, T] \times \Omega$), does not satisfy the initial condition $\tilde{\rho}(0) = \rho_0$, we do not know if \mathcal{B} admits a minimizer, and we do not know whether (21) holds.

To perform our analysis, let us fix $t_0 < t_1$ and a cut-off function $\zeta \in C_c^\infty([0, T])$ with $\zeta \equiv 1$ on $[t_0, t_1]$. Let us define

$$\begin{cases} \rho^{\eta, \delta}(t, x) := \rho(t + \zeta(t)\eta, x + \zeta(t)\delta), \\ v^{\eta, \delta}(t, x) := v(t + \zeta(t)\eta, x + \zeta(t)\delta)(1 + \zeta'(t)\eta) - \zeta'(t)\delta. \end{cases} \quad (22)$$

It is easy to check that the pair $(\rho^{\eta, \delta}, v^{\eta, \delta})$ satisfies the continuity equation together with the initial condition $\rho^{\eta, \delta}(0) = \rho_0$. Therefore, it is an admissible competitor in \mathcal{A} for any choice of (η, δ) . We may then consider the function

$$M : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}, \quad M(\eta, \delta) := \mathcal{A}(\rho^{\eta, \delta}, v^{\eta, \delta}).$$

The key point here is to show that M is smooth (actually, it would be enough to have $M \in C^{1,1}$).

Lemma 4.2 *The function $(\eta, \delta) \mapsto M(\eta, \delta)$ defined above is C^∞ .*

Proof We have

$$\mathcal{A}(\rho^{\eta, \delta}, v^{\eta, \delta}) = \int_0^T \int_{\mathbb{T}^d} \frac{1}{2} \rho^{\eta, \delta} |v^{\eta, \delta}|^2 dx dt + \int_0^T \int_{\mathbb{T}^d} G(\rho^{\eta, \delta}) dx dt + \int_{\mathbb{T}^d} \Psi(x) d\rho_T^{\eta, \delta}.$$

Since $\rho^{\eta, \delta}(T, x) = \rho(T, x)$, the last term does not depend on (η, δ) . For the other terms, we use the change of variable

$$(s, y) = (t + \zeta(t)\eta, x + \zeta(t)\delta)$$

which is a C^∞ diffeomorphism for small η . Then, we can write

$$\int_0^T \int_{\mathbb{T}^d} G(\rho^{\eta, \delta}(x, t)) dx dt = \int_0^T \int_{\mathbb{T}^d} G(\rho(y, s)) dy ds = \int_0^T \int_{\mathbb{T}^d} G(\rho(y, s)) K(\eta, \delta, s) dy ds,$$

where $K(\eta, \delta, s)$ is a smooth Jacobian factor (which does not depend on y since the change of variable is only a translation in space). Hence, this term depends smoothly on (η, δ) .

We also have

$$\begin{aligned} \int_0^T \int_{\mathbb{T}^d} \rho^{\eta, \delta} |v^{\eta, \delta}|^2 dx dt &= \int_0^T \int_{\mathbb{T}^d} \rho(s, y) |(1 + \eta \zeta'(t))v(s, y) - \delta \zeta'(t)|^2 dx dt \\ &= \int_0^T \int_{\mathbb{T}^d} \rho(s, y) |(1 + \eta \zeta'(t(\eta, s)))v(s, y) \\ &\quad - \delta \zeta'(t(\eta, s))|^2 K(\eta, \delta, s) dy ds, \end{aligned}$$

where $K(\eta, \delta, s)$ is the same Jacobian factor as before, and $t(\eta, s)$ is obtained by inverting, for fixed $\eta > 0$, the relation $s = t + \eta \zeta'(t)$, and is also a smooth map. Hence, this term is also smooth. \square

We can now apply the previous lemma to the estimate we need.

Proposition 4.3 *There exists a constant C , independent of (η, δ) , such that for $|\eta|, |\delta| \leq 1$, we have*

$$|M(\eta, \delta) - M(0, 0)| = |\mathcal{A}(\rho^{\eta, \delta}, v^{\eta, \delta}) - \mathcal{A}(\rho, v)| \leq C(|\eta|^2 + |\delta|^2).$$

Proof We just need to use Lemma 4.2 and the optimality of (ρ, v) . This means that M achieves its minimum at $(\eta, \delta) = (0, 0)$; therefore, its first derivative must vanish at $(0, 0)$ and we may conclude by a Taylor expansion. \square

With this result in mind, we may easily prove the following

Theorem 4.4 *If (ρ, v) is a solution to the primal problem $\min \mathcal{A}$, if $\Omega = \mathbb{T}^d$ and if J satisfies (20), then $J(\rho)$ satisfies, for every $t_0 < t_1$,*

$$\|J(\rho(\cdot + \eta, \cdot + \delta)) - J(\rho)\|_{L^2([t_0, t_1] \times \mathbb{T}^d)}^2 \leq C(|\eta|^2 + |\delta|^2)$$

(where the constant C depends on t_0, t_1 , and on the data), and hence is of class $H_{loc}^1([0, T] \times \mathbb{T}^d)$.

Proof Let us take a minimizing sequence (ϕ_n, p_n) for the dual problem, i.e., $\phi_n \in C^1$, $p_n = -\partial_t \phi_n + \frac{1}{2} |\nabla \phi_n|^2$ and

$$\mathcal{B}(\phi_n, p_n) \leq \inf_{(\phi, p) \in \mathcal{F}} \mathcal{B}(\phi, p) + \frac{1}{n}.$$

We use $\tilde{\rho} = \rho^{\eta, \delta}$ and $\tilde{v} = v^{\eta, \delta}$ as in the previous discussion. Using first the triangle inequality and then Lemma 4.1, we have (where the L^2 norm denotes the norm in $L^2((0, T) \times \mathbb{T}^d)$)

$$\begin{aligned} c_0 \|J(\rho^{\eta, \delta}) - J(\rho)\|_{L^2}^2 &\leq 2c_0 (\|J(\rho^{\eta, \delta}) - J_*(p_n)\|_{L^2}^2 + \|J(\rho) - J_*(p_n)\|_{L^2}^2) \\ &\leq 2(\mathcal{B}(\phi_n, p_n) + \mathcal{A}(\rho^{\eta, \delta}, v^{\eta, \delta}) + \mathcal{B}(\phi_n, p_n) + \mathcal{A}(\rho, v)), \end{aligned}$$

hence

$$\|J(\rho^{\eta, \delta}) - J(\rho)\|_{L^2}^2 \leq C(\mathcal{B}(\phi_n, p_n) + \mathcal{A}(\rho, v)) + C(|\eta|^2 + |\delta|^2) \leq \frac{C}{n} + C(|\eta|^2 + |\delta|^2).$$

Letting n go to infinity and restricting the L^2 norm to $[t_0, t_1] \times \mathbb{T}^d$, we get the claim. \square

Remark 3 If one restricts to the case $\eta = 0$, then it is also possible to use a cut-off function $\zeta \in C_c^\infty([0, T])$ with $\zeta(T) = 1$, as we only perform space translations. In this case, however, the final cost $\int_{\mathbb{T}^d} \Psi(x) d\rho_T^{\eta, \delta}$ depends on δ , and one needs to assume $\Psi \in C^{1,1}$ to prove $M \in C^{1,1}$. This allows to deduce H^1 regularity in space, local in time far from $t = 0$, i.e., $J(\rho) \in L^2_{loc}([0, T]; H^1(\mathbb{T}^d))$.

Remark 4 From $J(\rho) = J_*(p)$, the above regularity result on ρ can be translated into a corresponding regularity result on p .

Remark 5 How to handle the case of explicit dependance on x ? In this case, one should assume the existence of functions $J, J_* : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ such that (20) holds, for a uniform constant c_0 , for every x , in terms of $J(x, a)$ and $J_*(x, b)$. Then, Lemma 4.2 is no more evident and requires regularity of $x \mapsto G(x, \rho)$. Finally, the regularity which can be obtained is $J(x, \rho) \in H^1$, which can usually turn into regularity of ρ if J depends smoothly on x .

We stress that a finer analysis of the behavior at $t = T$ also allows to extend the above H^1 regularity result in space time till $t = T$, but needs extra tools (in particular defining a suitable extension of ρ for $t > T$). This is developed in [28].

Finally, we finish this section by underlining the regularity results in the density-constrained case ([16]): the same kind of strategy, but with many more technical

issues, which follow the same scheme as in [9] and [3], and the result is much weaker. Indeed, it is only possible to prove in this case $p \in L^2_{loc}((0, T); BV(\mathbb{T}^d))$ (exactly as in [3]). Even if very weak, this result is very important in what it gives higher integrability on p , which was a priori only supposed to be a measure, and this allows to get the necessary summability of the maximal function that we mentioned in Section 3.

5 Stochastic Control Variants

We now consider the case where the dynamics for the state of each player is governed by the controlled SDE $dX_t = \alpha_t dt + \sqrt{2} dW_t$, where the drift α_t is the agent's control and W_t is a standard Brownian motion; the goal of this section is to present (rather informally) some examples of MFG systems and their variational counterparts in various dynamic or static situations.

5.1 Dynamic MFG with Diffusion

Let us start with the finite horizon case where the goal of each agent starting from a position x at time 0, given the density of the other players ρ_t , is to solve the following stochastic control problem on the period $[0, T]$:

$$\inf_{\alpha} \left\{ \mathbb{E} \left[\int_0^T \left(\frac{1}{2} |\alpha_s|^2 + g(X_s, \rho_s(X_s)) \right) ds + \Psi(X_T) \right] : dX_s = \alpha_s ds + \sqrt{2} dW_s, X_0 = x \right\},$$

so that the value function

$$\phi(t, x) := \inf_{\alpha} \left\{ \mathbb{E} \left[\int_t^T \left(\frac{1}{2} |\alpha_s|^2 + g(X_s, \rho_s(X_s)) \right) ds + \Psi(X_T) \right] : dX_s = \alpha_s ds + \sqrt{2} dW_s, X_t = x \right\}$$

is governed by the following backward HJB equation

$$-\partial_t \phi - \Delta \phi + \frac{1}{2} |\nabla \phi|^2 = g(x, \rho), \quad \phi(T, .) = \Psi \quad (23)$$

and provided ϕ is smooth the optimal control in feedback form is given by $\alpha_t(x) = -\nabla \phi(t, x)$. Once we know this optimal drift and the initial distribution of players, the evolution of ρ_t is governed by the Fokker–Planck (or forward Kolmogorov equation):

$$\partial_t \rho - \Delta \rho - \nabla \cdot (\rho \nabla \phi) = 0, \quad \rho(0, .) = \rho_0. \quad (24)$$

The MFG system consists of the two equations (23)–(24) and corresponds to an equilibrium condition: the optimization of each player given their anticipation of

the density of the other players has to be consistent with the evolution of the players density resulting from their optimizing behavior. The fact that the MFG system (23)–(24) is related to an optimization problem (which is convex when g is non-decreasing, i.e., in the congested case) was first emphasized in the seminal works of Lasry and Lions [23, 24] and further analyzed by Cardaliaguet et al. in [15]. More precisely, defining G as before as the anti-derivative of g (extended by $+\infty$ on $(-\infty, 0)$) with respect to the second variable, and G^* its Legendre transform, the MFG system appears, at least formally (we refer to Cardaliaguet et al. in [15] for rigorous and detailed statements) as the optimality conditions for the two convex minimization problems in duality (for simplicity we consider again the periodic case $\Omega = \bar{\Omega}$):

$$\inf_{(\rho, w)} \{ \bar{\mathcal{A}}(\rho, w) : \partial_t \rho - \Delta \rho + \nabla \cdot w = 0, \rho(0, \cdot) = \rho_0 \} \quad (25)$$

where $\bar{\mathcal{A}}(\rho, w)$ is defined as in (8) and the dual

$$\inf_{\phi : \phi(T, \cdot) \leq \Psi} \int_0^T \int_{\Omega} G^* \left(x, -\partial_t \phi - \Delta \phi + \frac{1}{2} |\nabla \phi|^2 \right) dx dt - \int_{\Omega} \phi(0, x) d\rho_0(x). \quad (26)$$

5.2 Static MFGs with Noise

We now consider static situations with diffusion corresponding to different individual stochastic control problems for the players. It is now important to allow g to depend also on x (to avoid obvious situations such as a constant density being an equilibrium).

The Ergodic Problem

The ergodic MFG, first introduced in [22], corresponds to the case where each player aims at minimizing

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\int_0^T \left(\frac{1}{2} |\alpha_s|^2 + g(X_s, \rho(X_s)) \right) ds \right].$$

Given the measure ρ giving the (stationary) density of players, the value function satisfies the following HJB equation (again in the periodic setting for the sake of simplicity)

$$-\Delta \phi + \frac{1}{2} |\nabla \phi|^2 = \lambda + g(x, \rho) \text{ in } \Omega, \lambda \in \mathbb{R}. \quad (27)$$

At equilibrium ρ should be the invariant measure for the process corresponding to the optimal feedback $\alpha = -\nabla \phi$, i.e.,

$$-\Delta \rho - \nabla \cdot (\rho \nabla \phi) = 0 \text{ in } \Omega, \rho \geq 0, \int_{\Omega} \rho = 1. \quad (28)$$

Again, setting $G(x, \rho) = \int_0^\rho g(x, s)ds$ the MFG system (27)–(28) is formally the primal-dual system of optimality conditions for

$$\inf_{\rho, w} \left\{ \int_{\Omega} \left(\frac{|w|^2}{2\rho} + G(x, \rho) \right) dx, \quad \rho \in \mathcal{P}(\Omega), \quad -\Delta\rho + \nabla \cdot w = 0 \right\}$$

and its dual

$$\inf_{(\phi, \lambda)} \left\{ \int_{\Omega} G^* \left(x, (-\Delta\phi + \frac{1}{2}|\nabla\phi|^2 - \lambda)_+ \right) + \lambda \right\}.$$

We wish to mention in this framework the study which has been done in [27] about the same problem, with the constraint $\rho \leq 1$, where the (HJB) equation lets a pressure term appear on the saturated region $\{\rho = 1\}$.

The Exit Problem

Instead of the ergodic problem, it is also natural to look at the case where we are given a domain $\Omega \subset \mathbb{R}^d$ and starting from $x \in \Omega$, the player seeks to minimize

$$\mathbb{E} \left[\int_0^{\tau_x} \left(\frac{1}{2} |\alpha_s|^2 + g(X_s, \rho(X_s)) \right) ds + \Psi(X_{\tau_x}) \right]$$

where τ_x is the exit time from Ω , i.e., the first time at which $x + \int_0^t \alpha_s ds + \sqrt{2}W_t$ hits $\partial\Omega$ and Ψ is a given exit cost. This control problem corresponds of course to the Dirichlet problem for the stationary HJB equation

$$-\Delta\phi + \frac{1}{2}|\nabla\phi|^2 = g(x, \rho) \text{ in } \Omega, \quad \phi = \Psi \text{ on } \partial\Omega. \quad (29)$$

The MFG system then corresponds to the system formed of (29) coupled with (note that total mass is not fixed here)

$$-\Delta\rho - \nabla \cdot (\rho \nabla\phi) = 0 \text{ in } \Omega, \quad \rho \geq 0, \quad \rho = 0 \text{ on } \partial\Omega. \quad (30)$$

The optimality conditions for the convex minimization problem

$$\inf_{\phi : \phi|_{\partial\Omega} = \Psi} \left\{ \int_{\Omega} G^* \left(x, -\Delta\phi + \frac{1}{2}|\nabla\phi|^2 \right) \right\} \quad (31)$$

can then be (formally) written as follows: set $\rho := G^{*\prime} \left(x, -\Delta\phi + \frac{1}{2}|\nabla\phi|^2 \right)$ (the derivative is in the second variable, of course) then

$$\int_{\Omega} \rho(-\Delta u + \nabla\phi \cdot \nabla u) = 0, \quad \text{for all } u \in C^2(\overline{\Omega}), \quad u|_{\partial\Omega} = 0$$

i.e.,

$$-\Delta\rho - \nabla \cdot (\rho \nabla\phi) = 0, \quad \text{in } \Omega, \quad \rho|_{\partial\Omega} = 0$$

we thus recover (30), as for the HJB equation, since $\rho := G^{*\prime}(x, -\Delta\phi + \frac{1}{2}|\nabla\phi|^2)$, we have $g(x, \rho) = (-\Delta\phi + \frac{1}{2}|\nabla\phi|^2)$, i.e., ϕ satisfies (29). Finally, it can be easily checked that the equilibrium measure ρ can be obtained by solving the dual problem

$$\inf_{(\rho, w)} \left\{ \int_{\Omega} \left(\frac{|w|^2}{2\rho} + G(x, \rho) \right) dx + \int_{\partial\Omega} \Psi \frac{\partial \rho}{\partial n}, \quad \rho \geq 0, \quad -\Delta\rho + \nabla \cdot w = 0, \quad \rho = 0 \text{ on } \partial\Omega \right\}. \quad (32)$$

It is worth noting here that if $g(x, 0) \geq 0$, the problem is totally degenerate (and thus not really interesting); indeed in this case, G^* is minimal on \mathbb{R}_- , so every subsolution of the HJB equation $-\Delta\phi + \frac{1}{2}|\nabla\phi|^2 \leq 0$ coinciding with Ψ on $\partial\Omega$ solves (31) and the very degenerate density $\rho = 0$ is an equilibrium. This should come as no surprise since, as the mass is not fixed and the cost G is always minimal for $\rho = 0$, no player a priori wishes to enter this game! If on the contrary g is negative close to 0, the previous trivial situation $(\rho, w) = (0, 0)$ is in general not optimal for (32). This is the case, for instance, for logarithmic congestion functions.

The Discounted Infinite Horizon Problem

The last stationary situation we wish to discuss corresponds to the infinite horizon discounted criterion for the players:

$$\mathbb{E} \left[\int_0^\infty e^{-\lambda s} \left(\frac{1}{2} |\alpha_s|^2 + g(X_s, \rho(X_s)) \right) ds \right]$$

for a certain discount rate $\lambda > 0$. Such cases are particularly important for applications to macroeconomic dynamic models (see [1]) but as we shall see, they cannot be treated by a variational approach as the examples recalled above. Indeed the HJB equation for the value function reads

$$-\Delta\phi + \lambda\phi + \frac{1}{2}|\nabla\phi|^2 = g(x, \rho) \quad (33)$$

which is coupled with the same elliptic equation for the measure ρ as before, i.e., $-\Delta\rho - \nabla \cdot (\rho\nabla\phi) = 0$. Now it is quite clear that the corresponding system does not have a variational structure because the linear parts in the two equations are not adjoint: There is no $\lambda\rho$ term in the second equation!

One could add artificially this term, by considering a growth rate of the population and assuming that, by chance, this growth rate coincides with the discount rate λ , but this would lead to a different (and quite questionable) model... This example shows that there are of course limitations to the variational approach...

6 Numerical Simulations

6.1 Solving the MFG System by an Augmented Lagrangian Method

Our aim now is to explain how the variational problems for MFG systems recalled previously can be solved numerically by augmented Lagrangian methods and in particular the algorithm ALG2 of Fortin and Glowinski [20]. Such methods for the dynamical formulation of mass transport problems were used in the work of Benamou and Brenier [6]. Let us consider the deterministic evolutionary case as in Section 2 and let us rewrite the variational problem (6) (or some finite-element discretization of it) as

$$\inf_{(a,b,c,\phi)} \left\{ J(a, b, c) - \int_{\Omega} \phi(0, x) d\rho_0(x) : a = -\partial_t \phi, b = -\nabla \phi, c = \phi(T, .) \right\}$$

where

$$J(a, b, c) = \begin{cases} \int_0^T \int_{\Omega} G^*(a + \frac{1}{2}|b|^2) & \text{if } c \leq \Psi \\ +\infty & \text{otherwise} \end{cases}$$

which we may rewrite as an inf-sup (inf in (ϕ, a, b, c) and sup in (ρ, w, μ)) problem for the Lagrangian

$$\begin{aligned} \mathcal{L}(\phi, a, b, c, \rho, w, \mu) &= J(a, b, c) - \int_{\Omega} \phi(0, x) d\rho_0(x) \\ &\quad - \int_0^T \int_{\Omega} (\rho(t, x)(\partial_t \phi(t, x) + a(t, x)) + w(t, x) \cdot (\nabla \phi(t, x) \\ &\quad + b(t, x))) dx dt + \int_{\Omega} \mu(x)(\phi(T, x) - c(x)) dx \end{aligned}$$

or equivalently (see [20]) for the augmented Lagrangian

$$\begin{aligned} \mathcal{L}_r(\phi, a, b, c, \rho, w, \mu) &:= \mathcal{L}(\phi, a, b, c, \rho, w, \mu) \\ &\quad + \frac{r}{2} \left(\int_0^T \int_{\Omega} |(a + \partial_t \phi, b + \nabla \phi)|^2 dx dt + \int_{\Omega} (c(x) - \phi(T, x))^2 dx \right) \end{aligned}$$

where $r > 0$ (in practice in our simulations we will take $r = 1$). The augmented Lagrangian algorithm then consists, starting from an initial guess, in building inductively a sequence as follows: given $(\phi^k, a^k, b^k, c^k, \rho^k, w^k, \mu^k)$

Step 1: Find ϕ^{k+1} by minimizing $\mathcal{L}_r(., a^k, b^k, c^k, \rho^k, w^k, \mu^k)$; since this is a quadratic problem in $D_{t,x}\phi = (\partial_t \phi, \nabla \phi)$, it thus amounts to solve a Laplace equation (in the t and x variables) with suitable boundary conditions;

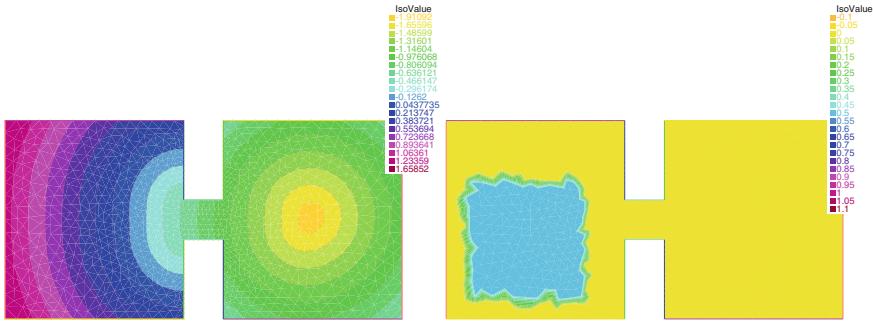
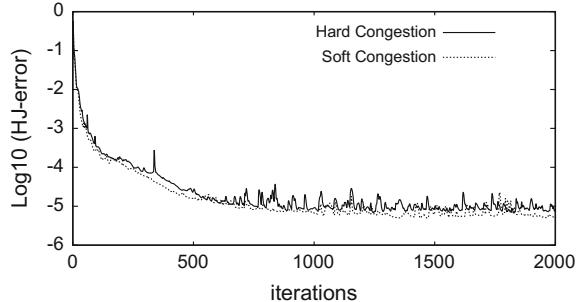


Fig. 1 On the left the final potential Ψ – on the right the initial crowd density

Fig. 2 Convergence of the ALG2 algorithm



Step 2: Find $(a^{k+1}, b^{k+1}, c^{k+1})$ by minimizing $\mathcal{L}_r(\phi^{k+1}, \dots, \rho^k, w^k, \mu^k)$; this consists in two pointwise proximal subproblems (one in (a, b) and one in c) which are in practice easy and quick to solve (see [7] for some details);

Step 3: Update the dual variables by the gradient ascent formula

$$(\rho^{k+1}, w^{k+1}) = (\rho^k - r(\partial_t \phi^{k+1} + a^{k+1}), w^k - r(\nabla \phi^{k+1} + b^{k+1}), \mu^k + r(\phi^{k+1}(T, \cdot) - c^{k+1})).$$

Note that this algorithm ensures that along the iterations the dual variables ρ^k, w^k remain such that $\rho^k \geq 0, w^k = 0$ whenever $\rho^k = 0$ so that one can define a velocity through $w^k = \rho^k v^k$ and the continuity equation

$$\partial_t \rho^k + \nabla \cdot (\rho^k v^k) = 0$$

is satisfied at each step. The algorithm can be adapted to diffusive cases as well. For evolutionary (respectively, stationary) diffusive cases, the relaxed variables become $a = -\partial_t \phi - \Delta \phi$ (resp. $a = -\Delta \phi$) and $b = -\nabla \phi$, the only significant modification then is that step 1 now involves an elliptic problem with the fourth-order operator $-\partial_{tt} + \Delta^2$ (resp. the bi-Laplacian Δ^2) for ϕ . See [2] for a recent ALG2 implementation in this case.

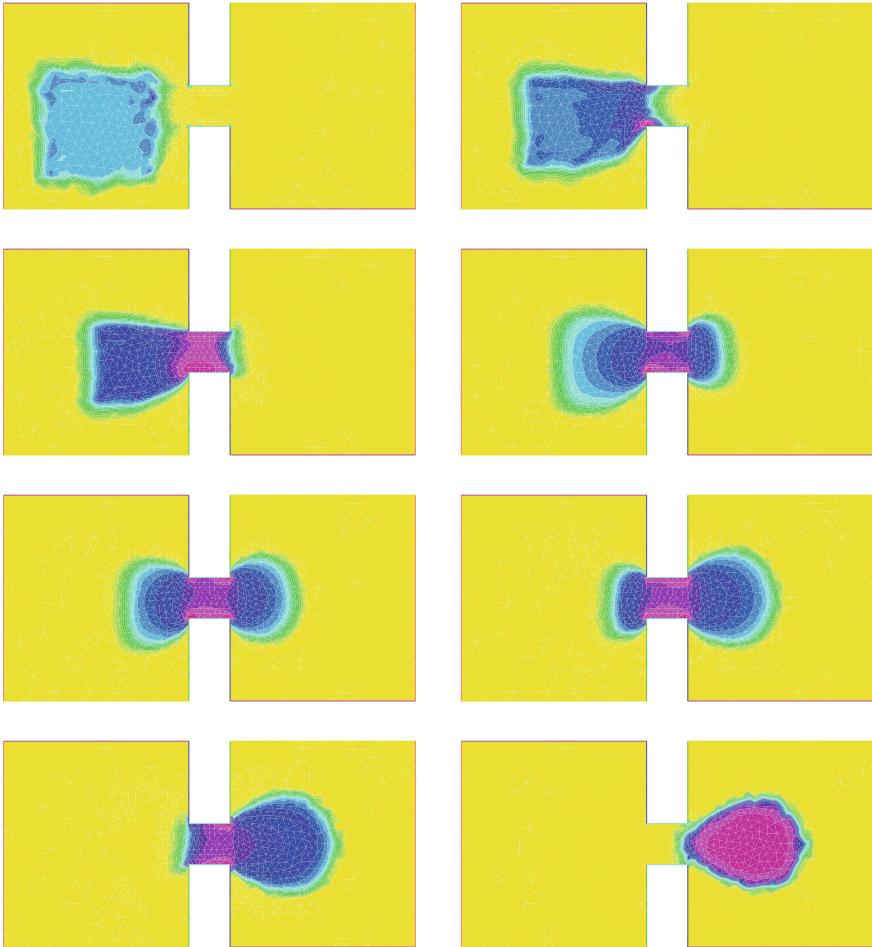


Fig. 3 Time evolution of the density. Soft congestion case $m = 6$. The color scale are the same as in Figure 1 right.

We present numerical results obtained with a FreeFem++ implementation adapted from [7].

6.2 Hard and Soft Congestion

We present simulations corresponding to the congestion models discussed at the end of Section 1.2. The macroscopic measure of the crowd density is ρ . Figure 1 shows the domain Ω made of two communicating rooms. The potential Ψ represented on

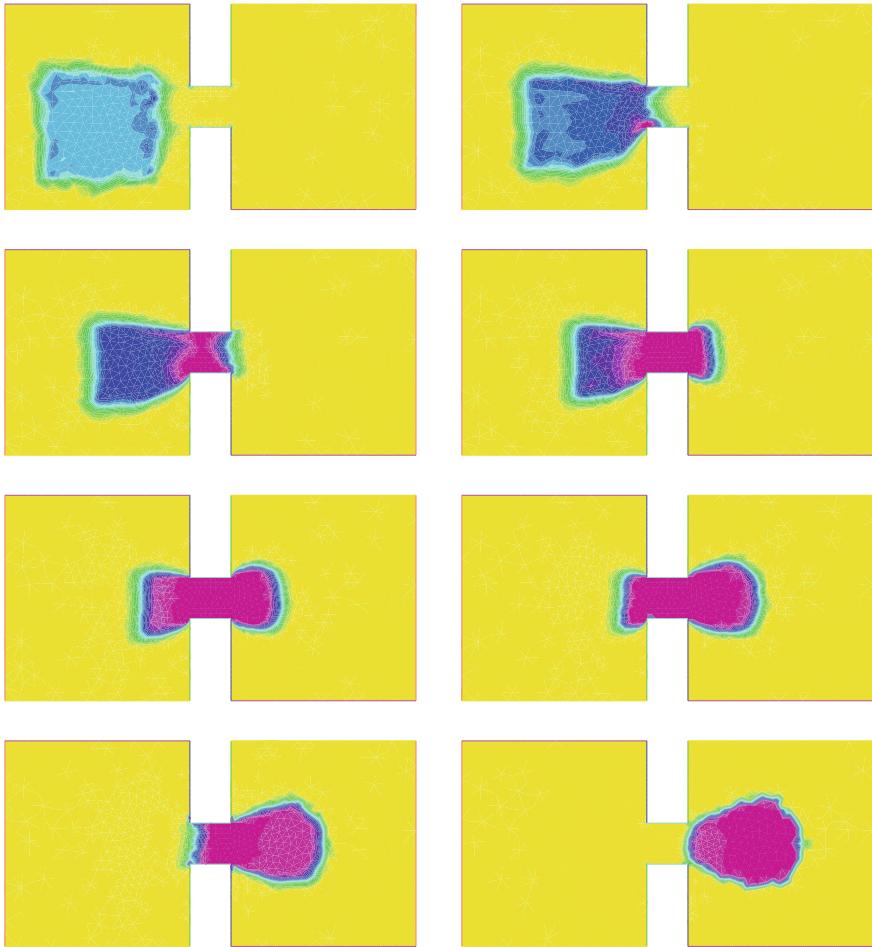


Fig. 4 Time evolution of the density. Hard congestion case. The color scale are the same as in Figure 1 right.

the left is a penalization which encourages agents to move from the first room to the second and gives target preferences within the rooms. On the right, we see the initial density (at time $t = 0$). The congestion is taken into account by the “cost” function G either in a “hard” way : $G(\rho) = 0$ if $\rho \in [0, 1]$ and $+\infty$ else or in a “soft” way $G(\rho) = \frac{\rho^m}{m}$ for $m > 1$. In the simulation below, we used $m = 6$.

Figure 2 shows the decrease in the L^2 residual of the Hamilton–Jacobi equation (the other equation in the coupled system, which is in this case a discretized continuity equation, is automatically satisfied after each ALG2 step).

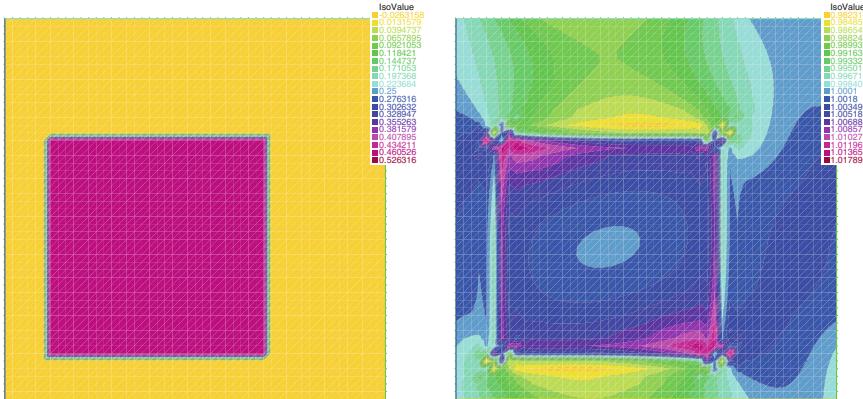


Fig. 5 On the left the potential ψ – on the right the invariant density

Figures 3 and 4 show time snapshot of the density in the hard, respectively, soft case. Agents move as expected to their optimal final position. As predicted by the analysis of ALG2, the density stays perfectly in $[0, 1]$.

6.3 Stationary Problem

Here, we show ALG2 simulations applied to the ergodic stationary models of Section 5.2 in a periodic domain. We use a linear form for the congestion $g(x, \rho) = \psi(x) \rho$ with a potential depending on x . Step 1 of ALG2 now involves a bi-laplacian operator. This is taken care of in FreeFem++ thanks to the recent addition of C1 conforming (HCT) finite elements.

The periodic square domain is discretized with a 50×50 grid, and this is the only parameter of the method! We show solutions for two different potentials ψ in Figures 5 and 6.

Figure 7 shows the decrease in the L^2 residual of the Hamilton–Jacobi equation (again, the other equation in the coupled system, $-\Delta \rho + \nabla \cdot w = 0$, is automatically satisfied after each ALG2 step).

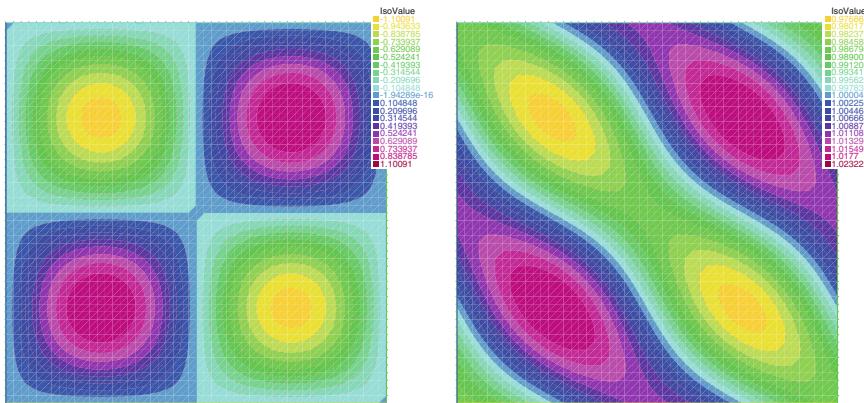
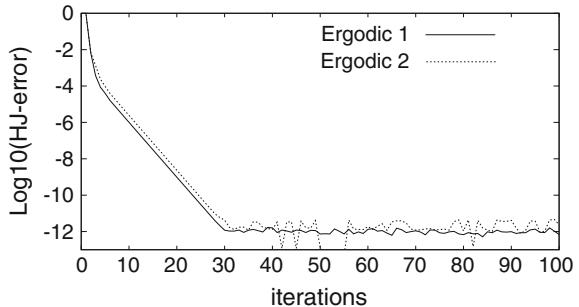


Fig. 6 On the left the potential ψ – on the right the invariant density

Fig. 7 Convergence of the ALG2 algorithm, for the two test cases in Figures 6 and 5.



Acknowledgements The authors acknowledge the support of the ANR project ISOTACE (ANR-12-MONU-0013). The third author also acknowledges the support of the iCODE project “Strategic Crowds,” funded by IDEX Paris-Saclay.

References

- Y. ACHDOU, F. J. BUERA, J.-M. LASRY, P.-L. LIONS, B. MOLL, Partial differential equation models in macroeconomics, *Phil. Trans. R. Soc. A* 372 (2014), 20130397.
- R. ANDREEV, Preconditioning the augmented Lagrangian method for instationary mean field games with diffusion. Available at <https://hal.archives-ouvertes.fr/hal-01301282>.
- L. AMBROSIO, A. FIGALLI, On the regularity of the pressure field of Brenier’s weak solutions to incompressible Euler equations, *Calc. Var. PDE*, 31 (2008) No. 4, 497-509.
- L. AMBROSIO, A. FIGALLI, Geodesics in the space of measure-preserving maps and plans, *Arch. Rational Mech. Anal.*, 194 (2009), 421-462.
- L. AMBROSIO, N. GIGLI, G. SAVARÉ, *Gradient flows in metric spaces and in the space of probability measures*, Lectures in Mathematics, Birkhäuser (2005).
- J.-D. BENAMOU, Y. BRENIER, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, *Numer. Math.*, 84 (2000), 375–393.
- J.-D. BENAMOU, G. CARLIER, Augmented Lagrangian Methods for Transport Optimization, Mean Field Games and Degenerate Elliptic Equations, *Journal of Optimization Theory and Applications* October 2015, Volume 167, Issue 1, pp 1-26.

8. G. BOUCHITTÉ, G. BUTTAZZO, New lower semicontinuity results for nonconvex functionals defined on measures, *Nonlinear Anal.*, 15 (1990), pp. 679–692.
9. Y. BRENIER, Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations, *Comm. Pure Appl. Math.*, 52 (1999) 4, 411–452.
10. G. BUTTAZZO, C. JIMENEZ, E. OUDET, An Optimization Problem for Mass Transportation with Congested Dynamics *SIAM J. Control Optim.* 48 (2010), 1961–1976.
11. P. CARDALIAGUET, Notes on Mean Field Games, available at <https://www.ceremade.dauphine.fr/~cardalia/MFG20130420.pdf>
12. P. CARDALIAGUET, Weak solutions for first order mean field games with local coupling, (2013), *preprint available at* <http://arxiv.org/abs/1305.7015>.
13. P. CARDALIAGUET, G. CARLIER, B. NAZARET, Geodesics for a class of distances in the space of probability measures, *Calc. Var. PDE*, 48 (2013), no. 2-3, 395–420.
14. P. CARDALIAGUET, J. GRABER, Mean field games systems of first order, *ESAIM: Contr. Opt. and Calc. Var.*, (2015), *to appear*.
15. P. CARDALIAGUET, J. GRABER, A. PORRETTA AND D. TONON, Second order mean field games with degenerate diffusion and local coupling, *Nonlinear Differ. Equ. Appl.*, 22 (2015), 1287–1317.
16. P. CARDALIAGUET, A. R. MÉSZÁROS, F. SANTAMBROGIO First order Mean Field Games with density constraints: pressure equals price. *SIAM Journal on Control and Optimization* 2016, Vol. 54, No. 5, pp. 2672–2709.
17. G. CARLIER, C. JIMENEZ, F. SANTAMBROGIO, Optimal transportation with traffic congestion and Wardrop equilibria, *SIAM J. Control Optim.* (47), 2008, 1330–1350.
18. G. CARLIER, F. SANTAMBROGIO, A continuous theory of traffic congestion and Wardrop equilibria, proceedings of *Optimization and stochastic methods for spatially distributed information*, St Petersburg, 2010. *Journal of Mathematical Sciences*, 181 (6), 792–804, 2012.
19. I. EKELAND, R. TEMAM, *Convex Analysis and Variational Problems*, Classics in Mathematics, Society for Industrial and Applied Mathematics (1999).
20. M. FORTIN, R. GLOWINSKI, *Augmented Lagrangian methods, Applications to the Numerical Solution of Boundary-Value Problems*, North-Holland (1983).
21. P.J. GRABER, Optimal control of first-order Hamilton-Jacobi equations with linearly bounded Hamiltonian, *Appl. Math. Optim.*, 70 (2014), no. 2, 185–224.
22. J.-M. LASRY, P.-L. LIONS, Jeux à champ moyen. I. Le cas stationnaire, *C. R. Math. Acad. Sci. Paris*, 343 (2006), No. 9, 619–625.
23. J.-M. LASRY, P.-L. LIONS, Jeux à champ moyen. II. Horizon fini et contrôle optimal, *C. R. Math. Acad. Sci. Paris*, 343 (2006), No. 10, 679–684.
24. J.-M. LASRY, P.-L. LIONS, Mean field games, *Jpn. J. Math.*, 2 (2007), no. 1, 229–260.
25. P.-L. LIONS, *Cours au Collège de France*, www.college-de-france.fr.
26. B. MAURY, A. ROUDNEFF-CHUPIN, F. SANTAMBROGIO A macroscopic crowd motion model of gradient flow type, *Math. Models and Methods in Appl. Sciences* Vol. 20, No. 10 (2010), 1787–1821.
27. A. R. MÉSZÁROS, F. J. SILVA A variational approach to second order mean field games with density constraints: the stationary case. *J. Math. Pures Appl.*, to appear.
28. A. PROSINSKI, F. SANTAMBROGIO Global-in-time regularity via duality for congestion-penalized Mean Field Games. Preprint available at cvgmt.sns.it.
29. F. SANTAMBROGIO, A modest proposal for MFG with density constraints, *Netw. Heterog. Media*, 7 (2012) No. 2, 337–347.
30. F. SANTAMBROGIO *Optimal Transport for Applied Mathematicians*, book, *Progress in Non-linear Differential Equations and Their Applications* 87, Birkhäuser Basel (2015).
31. F. SANTAMBROGIO Regularity via duality. Short lecture notes, available at <http://www.math.u-psud.fr/~santamb/LectureNotesDuality.pdf>
32. C. VILLANI *Topics in Optimal Transportation*. Graduate Studies in Mathematics, AMS, (2003).
33. J. G. WARDROP, Some theoretical aspects of road traffic research, *Proc. Inst. Civ. Eng.* 2 (1952), 325–378.

Sparse Control of Multiagent Systems

Mattia Bongini and Massimo Fornasier

Abstract In recent years, numerous studies have focused on the mathematical modeling of social dynamics, with *self-organization*, i.e., the autonomous pattern formation, as the main driving concept. Usually, first- or second-order models are employed to reproduce, at least qualitatively, certain global patterns (such as bird flocking, milling schools of fish, or queue formations in pedestrian flows, just to mention a few). It is, however, common experience that self-organization does not always spontaneously occur in a society. In this review chapter, we aim to describe the limitations of decentralized controls in restoring certain desired configurations and to address the question of whether it is possible to *externally* and *parsimoniously* influence the dynamics to reach a given outcome. More specifically, we address the issue of finding the sparsest control strategy for finite agent-based models in order to lead the dynamics optimally toward a desired pattern.

1 Introduction

The autonomous formation of patterns in multiagent dynamical systems is a fascinating phenomenon which has spawned an enormous wealth of interdisciplinary studies: from social and economic networks [6, 37], passing through cell aggregation and motility [13, 53, 55, 67], all the way to coordinated animal motion [17, 22, 27, 28, 30, 34, 62, 64, 65, 70, 75, 82] and crowd dynamics [2, 31, 36, 72]. Beyond biology and sociology, the principles of self-organization in multiagent systems are employed in engineering and information science to produce cheap, resilient, and efficient squadrons of autonomous machines to perform predefined tasks [3] and to render swarms of animals [69] and hair/fur textures in CGI animations [68]. The

M. Bongini · M. Fornasier (✉)

Technische Universität München, Fakultät Mathematik, Boltzmannstraße 3,
85748 Garching, Germany

e-mail: massimo.fornasier@ma.tum.de

M. Bongini

e-mail: mattia.bongini@ma.tum.de

scientific literature on the subject is vast and ever-growing: The interested reader may be addressed to [4, 19, 20, 76] and references therein for further insights on the topic.

A common feature of all those studies is that self-organization is the result of the superimposition of binary interactions between agents amplified by an accelerating feedback loop. This reinforcement process is necessary to give momentum to the multitude of feeble local interactions and to eventually let a global pattern appear. Typically, the strength of such interaction forces is a function of the “social distance” between agents: For instance, birds align with their closest neighbors [5] and people agree easier with those who already conform to their beliefs [51]. Some of the forces of the system may be of cohesive type; i.e., they tend to reduce the distance between agents: Whenever cohesive forces have a comparable strength at short and long range, we call these systems *heterophilious*; if, instead, there is a long-range bias, we speak of *homophilious* societies [60]. Heterophilious systems have a natural tendency to keep the trajectories of the agents inside a compact region and therefore to exhibit stable asymptotic profiles, modeling the autonomous emergence of global patterns. On the other hand, self-organization in homophilious societies can be accomplished only conditionally to sufficiently high levels of initial coherence that allow the cohesive forces to keep the dynamics compact [57]. Being such systems ubiquitous in real life (e.g., see [54]), it is legitimate to ask whether—in case of lost cohesion—additional forces acting on the agents of the system may restore stability and achieve pattern formation.

A first solution to facilitate self-organization is to consider *decentralized control strategies*: These consist in assuming that each agent, besides being subjected to forces induced in a feedback manner by the rest of the population, follows an individual strategy to coordinate with the other agents. However, as it was clarified in [11], even if we allow agents to self-steer toward consensus according to *additional decentralized feedback rules* computed with *local* information, their action results in general in a minor modification of the initial homophilious model, with no improvement in terms of promoting unconditional pattern formation. Hence, blindly insisting and believing on decentralized control is certainly fascinating, but rather wishful, as it does not secure self-organization.

Such additional forces may eventually be the result of an offline optimization among perfectly informed players: In this case, we fall into the realm of game theory [61, 78]. Games without an external regulator model situations where it is assumed that an automatic tendency to reach “correct” equilibria exists, like the stock market. However, also in this case, such an optimistic view of the dynamics is often frustrated by evidences of the convergence to suboptimal configurations [49], whence the need of an external figure controlling the evolution of the system.

For all these reasons, in the seminal papers [15, 16] *external*, controls with limited strength were considered to promote self-organization in multiagent systems. Notice that in such situations, *efficient* control strategies should target only few individuals of the population, instead of squandering resources on the entire group at once: Taking advantage of the mutual dependencies between the agents, they should trigger a ripple effect that would spread their influence to the whole system, thus indirectly

controlling the rest of the agents. The property of control strategies to target only a small fraction of the total population is known in the mathematical literature as *sparsity* [14, 40, 46]. The fundamental issue is the selection of the few agents to control: An effective criterion is to choose them as to maximize the decay rate of some Lyapunov functional associated with the stability of the desired pattern [25].

As a paradigmatic case study, let us consider *alignment models* [34, 51]: These are dissipative systems where imitation is the dominant feedback mechanism and in which the emerging pattern is a state where agents are fully aligned, also called *consensus*. For several of such models, it has been proved that consensus emergence can be guaranteed regardless of the initial conditions of the system only if the alignment forces are sufficiently strong at far distance, see [48, 50]; in case they are not, it is easy to provide counterexamples to the emergence of a consensus. If we were to use the criterion above to select a control strategy to steer the system to consensus, it would lead to a sparse control targeting at each instant only the agent farthest away from the mean consensus parameter. Surprisingly enough, for such systems, not only this strategy works for every initial condition, but the control of the instantaneous leaders of the dynamics is more convenient than controlling simultaneously all agents. Therefore if, on the one side, the homophilious character of a society plays against its compactness, on the other side, it may play at its advantage if we allow for sparse interventions to restore consensus.

The above results have more far-reaching potential as they can be extended to nondissipative systems as well, like the Cucker–Dong model of attraction and repulsion [33]. In this model, agents autonomously organize themselves in a *cohesive and collision-avoiding* configuration provided that the total energy is below a certain level. The sparse control strategy is able to raise this level considerably, and it is optimal in maximizing the convergence of the energy functional toward it. However in this case, due to the singular nonconservative forces in play, it may be seen that sparse controllability is in general conditional to the choice of the initial conditions, as opposed to the unconditional controllability of alignment models.

The essential scope of this review chapter is to describe in more detail the aforementioned mechanisms relating sparse controllability and pattern formation. We do so by condensing the results of the papers [8, 9, 11, 16], addressing the limitations of decentralized control strategies, the sparse controllability of alignment models, and the one of attraction repulsion models.

2 Self-organization in Dynamical Communication Networks

We start from the analysis of general properties of *alignment models*. Instances of these models are ubiquitous in nature since several species are able to interpret and instinctively reproduce certain manœuvres that they perceive (e.g., fleeing from a danger, searching for food, and performing defense tactics), see [39]. Such systems

may be seen as networks of agents with oriented information flow under possible link failure or creation and can be effectively represented by means of directed graphs with edges possibly switching in time.

A *directed graph* G on a set of nodes A_1, \dots, A_N is any subset of $\{A_1, \dots, A_N\}^2$. Each pair $(A, B) \in G$ is called an *edge from A to B*, and a *directed path from A to B* in G is a sequence of edges $(A, A_{i_1}), (A_{i_1}, A_{i_2}), \dots, (A_{i_k}, B) \in G$. The graph G is said to be *strongly connected* if for any pair A, B of distinct nodes there is a directed path from A to B and a directed path from B to A .

When studying under which conditions networks of agents are able to self-organize, it is usually not enough to know whether two nodes are connected: the strength of the interaction between them also matters. Hence, given a system of $N \in \mathbb{N}$ agents, for each pair of agents $i, j = 1, \dots, N$, we denote by $g_{ij}(t) \in \mathbb{R}_+$ the weight of the link connecting i with j : Clearly, if $g_{ij}(t) = 0$, i is not connected to j at time t . The value $g_{ij}(t)$ can be seen as the relative intensity of the information exchange flowing from agent i to agent j at time $t \geq 0$. We shall assume for the moment that each weight function $g_{ij} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is piecewise continuous.

The weights $g_{ij}(\cdot)$ naturally induce a directed graph structure on the set of agents: We define, for any $\varepsilon \geq 0$ and $t \geq 0$, the graph $G_\varepsilon(t)$ as

$$G_\varepsilon(t) \triangleq \{(i, j) \in \{1, \dots, N\}^2 : g_{ij}(t) > \varepsilon\}.$$

The *adjacency matrix* $G_0(t)$ is the set of pairs (i, j) for which the communication channel from i to j is active at time t .

As a prototypical example of a multiagent system and to quantitatively illustrate the concept of self-organization, we introduce *alignment models*: If we denote by $\{v_1, \dots, v_N\} \subset \mathbb{R}^d$ the states of the N agents of our systems, then the instantaneous evolution of the state $v_i(t)$ of agent i at time t is given by

$$\dot{v}_i(t) = \sum_{j=1}^N g_{ij}(t)(v_j(t) - v_i(t)), \quad i = 1, \dots, N. \quad (1)$$

The meaning of the above system of differential equations is the following: At each instant $t \geq 0$, the state $v_i(t)$ of agent i tends to the state $v_j(t)$ of agent j with a speed that depends on the strength of the information exchange $g_{ij}(t)$. Since (1) is a system of ODEs with possibly discontinuous coefficients, we need for it a proper notion of solution.

Definition 1 Let $\{I_k\}_{k \in \mathbb{N}}$ denote a countable family of open intervals such that all the functions g_{ij} are continuous on every I_k and $\cup_{k \in \mathbb{N}} I_k = \mathbb{R}_+$. Given $v^0 = (v_1^0, \dots, v_N^0) \in \mathbb{R}^{dN}$, we say that the curve $v = (v_1, \dots, v_N) : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN}$ is a *solution* of (1) with initial datum v^0 if

- (i) $v(0) = v^0$;
- (ii) for every $i = 1, \dots, N$ and $k \in \mathbb{N}$, v_i satisfies (1) on I_k .

The notion of self-organization that we are considering for system (1) is that of *consensus* or *flocking*, which is the situation where the state variables of the agents asymptotically coincide.

Definition 2 (Consensus for system (1)) Let $v : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN}$ denote a solution of (1) with initial datum v^0 . We say that $v(\cdot)$ converges to consensus if there exists a $v^\infty \in \mathbb{R}^d$ such that, for every $i = 1, \dots, N$, it holds

$$\lim_{t \rightarrow +\infty} \|v_i(t) - v^\infty\|_{\ell_2^d} = 0.$$

The value v^∞ is called the *consensus state*.

In the definition above, $\|\cdot\|_{\ell_2^d}$ stands for the Euclidean norm on \mathbb{R}^d . The subscript ℓ_2^d shall often be omitted whenever clear from context.

Roughly speaking, a system of agents satisfying (1) converges to consensus regardless of the initial condition v^0 provided that the underlying communication graph is “sufficiently connected.” With this we mean that each node must possess, over some dense collection of time intervals, a strong enough communication path to every other nodes in the network. This intuitive idea is made precise in the following result, whose proof can be found in [50]. A similar answer for discrete-time systems was also provided in [59]

Theorem 1 *Let $v : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN}$ be a solution of (1) with initial datum v^0 . Suppose that there exists an $\varepsilon > 0$ and a strongly connected directed graph G on the set of agents on which the system spends an infinite amount of time, i.e.,*

$$\mathcal{L}^1(\{t \geq 0 : G_\varepsilon(t) = G\}) = +\infty.$$

Then $v(\cdot)$ converges to consensus with consensus state v^∞ belonging to the convex hull of $\{v_1^0, \dots, v_N^0\}$.

The above result is closely related, for instance, to [60, Theorem 2.3], which requires a stronger connectivity of the network of agents (the quantity η_A in [60, Equation (2.5)]) but also gives an explicit rate for the convergence toward v^∞ (see [60, Equation (2.6b)]).

Theorem 1 also says that without further hypotheses on the interaction weights g_{ij} , the value of v^∞ is rather an emergent property of the global dynamics of system (1) than a mere function of the initial datum v^0 . Nonetheless, it is relatively simple to identify assumptions on g_{ij} for which the latter is true. For example, from a trivial computation follows

$$\frac{1}{N} \sum_{i=1}^N \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^N g_{ij}(t) - \sum_{i=1}^N g_{ji}(t) \right) v_j(t).$$

Hence, if for every $t \geq 0$ the weight matrix $(g_{ij}(t))_{i,j=1}^N$ has the property that $\sum_{i=1}^N g_{ij}(t) = \sum_{i=1}^N g_{ji}(t)$ for every $j = 1, \dots, N$, then the average

$$\bar{v}(t) \triangleq \frac{1}{N} \sum_{i=1}^N v_i(t) \quad (2)$$

is an invariant of the dynamics. This implies that $v^\infty = \bar{v}(0)$ holds; i.e., the consensus state is only a function of the initial datum v^0 .

3 Consensus Emergence in Alignment Models

In this section, we shall see that the assumptions of Theorem 1 can actually be very restrictive and seldom met when dealing with specific instances of alignment models.

3.1 Some Classic Examples of Alignment Models

A general principle in opinion formation is the *conformity bias*; i.e., agents weight more opinions that already conform to their beliefs. This can, actually, be extended to coordination in general, since intuitively it is easier to coordinate with “near” agents than “far away” ones. Formally, this is equivalent to asking that the weights g_{ij} are a nonincreasing function of the distance between the states of the agents, i.e.,

$$g_{ij}(t) = a(\|v_i(t) - v_j(t)\|), \quad (3)$$

where $a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a nonincreasing *interaction kernel*. Notice that (3) trivially implies the invariance of the mean \bar{v} (given by (2)) and that $v^\infty = \bar{v}(0)$, if it exists.

Several classic opinion formation models combine conformity bias with alignment. In the DW model, see [80], two random agents i and j update their opinions v_i and v_j to $1/2(v_i + v_j)$, provided they originally satisfy $\|v_i - v_j\| \leq R$, where $R > 0$ is fixed *a priori*. Instead, in the popular *bounded confidence* model of Hegselmann and Krause [51], opinions evolve according to the dynamics (1) where the function a has the form

$$a(r) = \chi_{[0, R]}(r) \triangleq \begin{cases} 1 & \text{if } r \in [0, R], \\ 0 & \text{otherwise,} \end{cases}$$

for some fixed *confidence radius* $R > 0$. The dynamics is thus given by the system of ODEs

$$\dot{v}_i(t) = \frac{1}{|\Lambda_R(t, i)|} \sum_{j=1}^N \chi_{[0, R]}(\|v_i(t) - v_j(t)\|)(v_j(t) - v_i(t)), \quad i = 1, \dots, N, \quad (4)$$

where we have set

$$\Lambda_R(t, i) \triangleq \{j \in \{1, \dots, N\} : \|x_i(t) - x_j(t)\| \leq R\}, \quad (5)$$

and $|\Lambda_R(t, i)|$ stands for its cardinality. It is straightforward to design an instance of this model not fulfilling the hypothesis of Theorem 1. Indeed, consider a group of $N = 2$ agents in dimension $d = 1$ with initial conditions $v_1(0) = -R$ and $v_2(0) = R$. Since $g_{12}(0) = g_{21}(0) = 0$, it follows that $G_\varepsilon(t) = \emptyset$ for all $t \geq 0$ and for all $\varepsilon \geq 0$.

Second-order models are necessary whenever we want to describe the dynamics of physical agents, like flocks of birds, herds of quadrupeds, schools of fish, and colonies of bacteria, where individuals are considered aligned whenever they move in the same direction, regardless of their position. Since in such cases it is necessary to perceive the velocities of the others in order to align, to describe the motion of the agents we need the pair position-velocity (x, v) , but this time only the velocity variable v is the *consensus parameter*.

One of the first of such models, named *Vicsek model* in honor of one of its fathers, was introduced in [77]. Very much in the spirit of (4), it postulates that the evolution of the spatial coordinate x_i and that of the orientation $\theta_i \in [0, 2\pi]$ in the plane \mathbb{R}^2 of the i th agent follow the law of motion given by

$$\begin{cases} \dot{x}_i(t) = v_i(t) = \hat{v} \begin{pmatrix} \cos(\theta_i(t)) \\ \sin(\theta_i(t)) \end{pmatrix}, \\ \dot{\theta}_i(t) = \frac{1}{|\Lambda_R(t, i)|} \sum_{j=1}^N \chi_{[0, R]}(\|x_i(t) - x_j(t)\|) (\theta_j(t) - \theta_i(t)), \end{cases} \quad i = 1, \dots, N, \quad (6)$$

where $\hat{v} > 0$ denotes the constant modulus of $v_i(t)$. In this model, the orientation of the consensus parameter v_i is adjusted with respect to the other agents according to a weighted average of the differences $\theta_j - \theta_i$. The influence of the j th agent on the dynamics of the i th one is a function of the (physical or social) distance between the two agents: If this distance is less than R , the agents interact by appearing in the computation of the respective future orientation (Figure 1).

In [34], the authors proposed a possible extension of system (6) to dimensions $d > 2$ as follows

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{|\Lambda_R(t, i)|} \sum_{j=1}^N \chi_{[0, R]}(\|x_i(t) - x_j(t)\|) (v_j(t) - v_i(t)), \end{cases} \quad i = 1, \dots, N.$$

The substitution of the function $\chi_{[0, R]}$ with a strictly positive kernel $a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ let us drop the highly irregular and nonsymmetric normalizing factor $|\Lambda_R(t, i)|$ in favor of a simple N and lead to the system

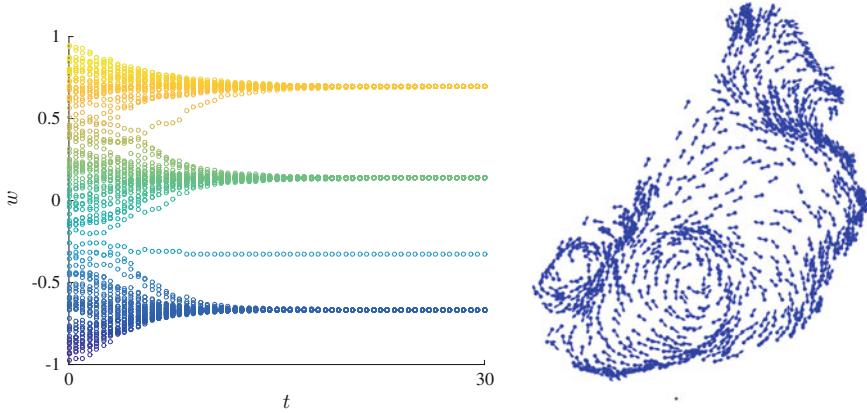


Fig. 1 On the left: a typical evolution of the Hegselmann–Krause model. On the right: mill patterns in the Vicsek model. (Kind courtesy of G. Albi)

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|) (v_j(t) - v_i(t)), \end{cases} \quad i = 1, \dots, N. \quad (7)$$

Notice that the equation governing the evolution of v_i has the same form as (1), and since now the weights g_{ij} are symmetric (i.e., $g_{ij} = g_{ji}$ for all $i, j = 1, \dots, N$), then \bar{v} is a conserved quantity.

An example of a system of the form (7) is the influential model of Cucker and Smale, introduced in [34], in which the function a is

$$a(r) \triangleq \frac{H}{(\sigma^2 + r^2)^\beta}, \quad (8)$$

where $H > 0$, $\sigma > 0$, and $\beta \geq 0$ are constants accounting for the social properties of the group. Systems like (7) are usually referred to as *Cucker–Smale systems* due to the influence of their work, as can be witnessed by the wealth of the literature focusing on their model, see for instance [1, 18, 38, 47, 66, 71].

3.2 Pattern Formation for the Cucker–Smale Model

We now focus on consensus emergence for system (7). In the following, we shall consider a kernel $a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ which is decreasing, strictly positive, bounded, and Lipschitz continuous.

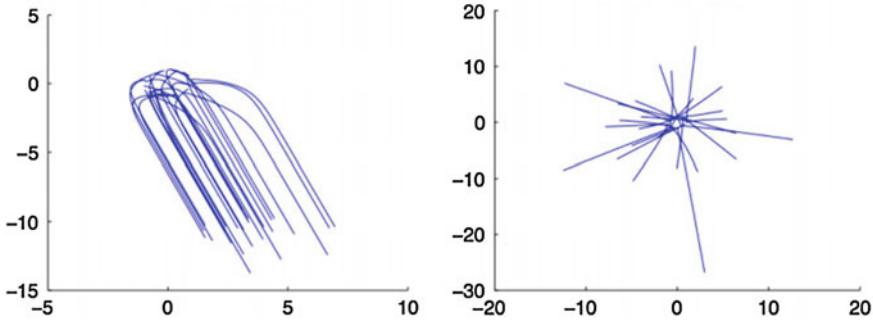


Fig. 2 Consensus behavior of a Cucker–Smale system. On the left: agents align with the mean velocity. On the right: agents fail to reach consensus.

As already noticed, in second-order models, alignment means that all agents move with the same velocity, but *not necessarily* are in the same position. Therefore, Definition 2 of consensus applies here on the v_i variables only (Figure 2).

Definition 3 (Consensus for system (7)) We say that a solution

$$(x, v) = (x_1, \dots, x_N, v_1, \dots, v_N) : \mathbb{R}_+ \rightarrow \mathbb{R}^{2dN}$$

of system (7) *tends to consensus* if the consensus parameter vectors v_i tend to the mean \bar{v} , i.e.,

$$\lim_{t \rightarrow +\infty} \|v_i(t) - \bar{v}(t)\| = 0 \quad \text{for every } i = 1, \dots, N.$$

The following result is an easy corollary of Theorem 1.

Corollary 1 Let $(x(\cdot), v(\cdot))$ be a solution of system (7), where the interaction kernel a is decreasing and strictly positive. Suppose that there exists $R > 0$ for which it holds

$$\mathcal{L}^1(\{t \geq 0 : \|x_i(t) - x_j(t)\| \leq R \text{ for all } i, j = 1, \dots, N\}) = +\infty.$$

Then $(x(\cdot), v(\cdot))$ converges to consensus.

Proof Since a is decreasing and strictly positive, from the initial assumptions follows

$$g_{ij}(t) = \frac{1}{N} a(\|x_i(t) - x_j(t)\|) \geq \frac{a(R)}{N} > 0,$$

for every $t \geq 0$ for which $\|x_i(t) - x_j(t)\| \leq R$ holds for every $i, j = 1, \dots, N$. Therefore, the condition $\|x_i(t) - x_j(t)\| \leq R$ for every $i, j = 1, \dots, N$ implies $G_{a(R)/N}(t) = \{1, \dots, N\}^2$, which yields

$$\mathcal{L}^1(\{t \geq 0 : G_{a(R)/N}(t) = \{1, \dots, N\}^2\}) = +\infty,$$

The statement then follows from Theorem 1 for the choice $\varepsilon = a(R)/N$.

Unfortunately, the result above has the serious flaw that it cannot be invoked directly to infer convergence to consensus, since establishing a uniform bound in time for the distances of the agents is very difficult, even for smooth kernels like (8). Intuitively, consider the case where the interaction strength is too weak and the agents too dispersed in space to let the velocities v_i align. In this case, nothing prevents the distances $\|x_i - x_j\|$ to grow indefinitely, violating the hypothesis of Corollary 1. Hence, in order to obtain more satisfactory consensus results, we need to follow approaches that take into account the extra information at our disposal, which are the strength of the interaction and the initial configuration of the system.

Originally, this problem was studied in [34, 35] borrowing several tools from spectral graph theory, see as a reference [23]. Indeed, system (7) can be rewritten in the following compact form

$$\begin{cases} \dot{x}(t) = v(t), \\ \dot{v}(t) = L(x(t))v(t), \end{cases} \quad (9)$$

where $L(x(t))$ is the Laplacian¹ of the matrix $(a(\|x_i(t) - x_j(t)\|)/N)_{i,j=1}^N$, which is a function of $x(t)$. Being the Laplacian of a positive-definite, symmetric matrix, $L(x(t))$ encodes plenty of information regarding the adjacency matrix $G_0(t)$ of the system, see [58]. In particular, the second smallest eigenvalue $\lambda_2(t)$ of $L(x(t))$, called the *Fiedler's number* of $G_0(t)$, is deeply linked with consensus emergence: Provided that a sufficiently strong bound from below of $\lambda_2(t)$ is available, the system converges to consensus.

To establish under which conditions we have convergence to consensus, we shall follow a different approach. The advantage of it is that it can be employed also to study the issue of the controllability of several multiagent systems (see Section 5).

3.3 The Consensus Region

A natural strategy to improve Corollary 1 would be to look for quantities which are invariant with respect to \bar{v} , since it is conserved in systems like (7).

Definition 4 The symmetric bilinear form $B : \mathbb{R}^{dN} \times \mathbb{R}^{dN} \rightarrow \mathbb{R}$ is defined, for any $v, w \in \mathbb{R}^{dN}$, as

¹Given a real $N \times N$ matrix $A = (a_{ij})_{i,j=1}^N$ and $v \in \mathbb{R}^{dN}$, we denote by Av the action of A on \mathbb{R}^{dN} by mapping v to $(a_{i1}v_1 + \dots + a_{iN}v_N)_{i=1}^N$. Given a nonnegative symmetric $N \times N$ matrix $A = (a_{ij})_{i,j=1}^N$, the *Laplacian* L of A is defined by $L = D - A$, with $D = \text{diag}(d_1, \dots, d_N)$ and $d_k = \sum_{j=1}^N a_{kj}$.

$$B(v, w) \triangleq \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N (v_i - v_j) \cdot (w_i - w_j),$$

where \cdot denotes the usual scalar product on \mathbb{R}^d .

Remark 1 It is trivial to prove that

$$B(v, w) = \frac{1}{N} \sum_{i=1}^N (v_i \cdot w_i) - \bar{v} \cdot \bar{w}, \quad (10)$$

where \bar{v} stands for the average of the elements of the vector $v = (v_1, \dots, v_N)$ given by (2). From this representation of B follows easily that the two spaces

$$\mathcal{V}_f \triangleq \left\{ v \in \mathbb{R}^{dN} : v_1 = \dots = v_N \right\} \quad \text{and} \quad \mathcal{V}_\perp \triangleq \left\{ v \in \mathbb{R}^{dN} : \sum_{i=1}^N v_i = 0 \right\},$$

are perpendicular with respect to the scalar product B , i.e., $\mathbb{R}^{dN} = \mathcal{V}_f \oplus \mathcal{V}_\perp$. This means that every $v \in \mathbb{R}^{dN}$ can be written uniquely as $v = v^f + v^\perp$, where $v^f \in \mathcal{V}_f$ and $v^\perp \in \mathcal{V}_\perp$. A closer inspection reveals that it holds $v_i^f = \bar{v}$ and $v_i^\perp = v_i - \bar{v}$ for every $i = 1, \dots, N$. Notice that since $v^\perp \in \mathcal{V}_\perp$, for any vector $w \in \mathbb{R}^d$, it holds

$$\sum_{i=1}^N (v_i^\perp \cdot w) = \left(\sum_{i=1}^N v_i^\perp \right) \cdot w = 0. \quad (11)$$

Since for every $v, w \in \mathbb{R}^{dN}$ we have $B(v^f, w) = 0 = B(v, w^f)$, it holds

$$B(v, w) = B(v^\perp, w) = B(v, w^\perp) = B(v^\perp, w^\perp).$$

This means that B distinguishes two vectors modulo their projection on \mathcal{V}_f . Moreover, from (10) immediately follows that B restricted to $\mathcal{V}_\perp \times \mathcal{V}_\perp$ coincides, up to a factor $1/N$, with the usual scalar product on \mathbb{R}^{dN} .

Remark 2 (Consensus manifold) Notice that whenever the initial datum (x^0, v^0) belongs to the set $\mathbb{R}^{dN} \times \mathcal{V}_f$, the right-hand size of \dot{v}_i in (7) is 0; hence, the equality $v_1(t) = \dots = v_N(t)$ is satisfied for all $t \geq 0$ and the system is already in consensus. For this reason, the set $\mathbb{R}^{dN} \times \mathcal{V}_f$ is called the *consensus manifold*.

The bilinear form B can be used to characterize consensus emergence for solutions $(x(\cdot), v(\cdot))$ of system (7) by setting

$$X(t) \triangleq B(x(t), x(t)) \quad \text{and} \quad V(t) \triangleq B(v(t), v(t)).$$

The functionals X and V provide a description of consensus by measuring the spread, both in positions and velocities, of the trajectories of the solution $(x(\cdot), v(\cdot))$, as the following trivial result shows.

Proposition 1 *The following statements are equivalent:*

1. $\lim_{t \rightarrow +\infty} \|v_i(t) - \bar{v}(t)\| = 0$ for every $i = 1, \dots, N$;
2. $\lim_{t \rightarrow +\infty} v_i^\perp(t) = 0$ for every $i = 1, \dots, N$;
3. $\lim_{t \rightarrow +\infty} V(t) = 0$.

The following Lemma shows that V is a Lyapunov functional for system (7).

Lemma 1 ([16, Lemma 1]) *Let $(x(\cdot), v(\cdot))$ be a solution of system (7). Then for every $t \geq 0$ it holds*

$$\frac{d}{dt} V(t) \leq -2a(\sqrt{2N}X(t))V(t). \quad (12)$$

Therefore, V is decreasing.

By means of the quantities X and V , we can provide a sufficient condition for consensus emergence for solutions of system (7).

Theorem 2 ([48, Theorem 3.1]) *Let $(x^0, v^0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ and set $X_0 \triangleq B(x^0, x^0)$ and $V_0 \triangleq B(v^0, v^0)$. If the following inequality is satisfied*

$$\int_{\sqrt{X_0}}^{+\infty} a(\sqrt{2Nr}) dr \geq \sqrt{V_0}, \quad (13)$$

then the solution of (7) with initial datum (x^0, v^0) tends to consensus.

The inequality (13) defines a region in the space (X_0, V_0) of initial conditions for which the balance between X_0 , V_0 and the kernel a is such that the system tends to consensus autonomously.

Definition 5 (Consensus region) We call *consensus region* the set of points $(X_0, V_0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ satisfying (13).

The size of the consensus region gives an estimate of how large the basin of attraction of the consensus manifold $\mathbb{R}^{dN} \times \mathcal{V}_f$ is. If the rate of communication function a is integrable, i.e., far distant agents are only weakly influencing the dynamics, then such a region is essentially bounded, and actually not all initial conditions will realize self-organization, as the following example shows.

Example 1 ([34, Proposition 5]) Consider $N = 2$ agents in dimension $d = 1$ subject to system (7) with interaction kernel given by (8) with $H = 1/2$, $\sigma = 1$, and $\beta = 1$. If we denote by $(x_1(\cdot), v_1(\cdot))$ and $(x_2(\cdot), v_2(\cdot))$ the trajectories of the two agents, it

is easy to show that the evolution of the relative main state $x(t) \stackrel{\Delta}{=} x_1(t) - x_2(t)$ and that of the relative consensus state $v(t) \stackrel{\Delta}{=} v_1(t) - v_2(t)$ are given for every $t \geq 0$ by

$$\begin{cases} \dot{x}(t) = v(t), \\ \dot{v}(t) = -\frac{v(t)}{1+x(t)^2}, \end{cases} \quad (14)$$

with initial conditions $x(0) = x^0$ and $v(0) = v^0$ (without loss of generality, we may assume that $x^0, v^0 > 0$). An explicit solution of the above system can be easily derived by means of direct integration:

$$v(t) - v^0 = -\arctan x(t) + \arctan x^0.$$

Condition (13) in this case reads $\pi/2 - \arctan x^0 \geq v^0$. Hence, suppose (13) is violated, i.e., $\arctan x^0 + v^0 > \pi/2$. This means $\arctan x^0 + v^0 \geq \pi/2 + \varepsilon$ for some $\varepsilon > 0$, which implies

$$|v(t)| = |-\arctan x(t) + \arctan x^0 + v^0| \geq \left| -\arctan x(t) + \frac{\pi}{2} + \varepsilon \right| > \varepsilon$$

for every $t \geq 0$. Therefore, the solution of system (14) with initial datum (x^0, v^0) satisfying $\arctan x^0 + v^0 > \pi/2$ does not converge to consensus, since otherwise we would have $v(t) \rightarrow 0$ for $t \rightarrow +\infty$.

Remark 3 Notice that if $\int_{\delta}^{+\infty} a(r) dr$ diverges for every $\delta \geq 0$, then the consensus region coincides with the entire space $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$. In other words, in this case, the interaction force between the agents is so strong that the system will reach consensus no matter what the initial conditions are.

As the following example shows, there may be initial configurations from which the system can reach consensus automatically even if condition (13) is not satisfied.

Example 2 Consider an instance of the Cucker–Smale system (7) without control in dimension $d = 1$ with $N = 2$ agents, where the interaction function $a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is of the form

$$a(r) = \begin{cases} M & \text{if } r \leq R, \\ f(r) & \text{if } r \geq R, \end{cases}$$

for some given $R > 0$ and $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ positive continuous function satisfying

$$f(R) = M \quad \text{and} \quad \int_R^{+\infty} f(r) dr = \varepsilon < +\infty.$$

The constant $M > 0$ is to be properly chosen later on. Assume that the initial state and consensus parameters of the two agents are $(x_1^0, v_1^0) = (-R/2, v^0)$ and $(x_2^0, v_2^0) = (R/2, -v^0)$, respectively, for some $v^0 > \varepsilon/2$.

Due to the nature of the situation, it is fairly easy to check whether condition (13) of Theorem 2 is satisfied or not. Indeed, we have $X(0) = R^2/4$ and $V(0) = (v^0)^2$, and by the particular form of a , after a change of variables, the computation below follows

$$\int_{\frac{R}{2}}^{+\infty} a(2r) dr = \frac{1}{2} \int_R^{+\infty} a(r) dr = \frac{1}{2} \int_R^{+\infty} f(r) dr = \frac{\varepsilon}{2}.$$

Therefore, at time $t = 0$, we are not in the consensus region given by (13), since

$$\int_{\sqrt{X(0)}}^{+\infty} a(\sqrt{4r}) dr = \frac{\varepsilon}{2} < v^0 = \sqrt{V(0)}.$$

We now show that there exists a time $T > 0$ such that

$$\int_{\sqrt{X(T)}}^{+\infty} a(\sqrt{4r}) dr \geq \sqrt{V(T)}, \quad (15)$$

i.e., the system enters the consensus region autonomously at time T .

To do so, we first compute a lower bound for the integral. Notice that since we are considering a Cucker–Smale system with mean consensus parameter $\bar{v} = 0$, the speeds $|v_1(t)|$ and $|v_2(t)|$ are decreasing by Lemma 1. Therefore, we can estimate from above the time until $|x_1(t) - x_2(t)| \leq R$ holds by $T^* \triangleq R/2v^0$ (since the agents are moving on the real line in opposite directions). Hence, $X(t) \leq X(0) = R^2/4$ for every $t \in [0, T^*]$, which yields the following lower bound

$$\int_{\sqrt{X(t)}}^{+\infty} a(\sqrt{4r}) dr \geq \int_{\sqrt{X(0)}}^{+\infty} a(\sqrt{4r}) dr = \frac{\varepsilon}{2}$$

valid for any $t \leq T^*$.

We now compute an upper bound for the functional $\sqrt{V(t)}$ for $t \in [0, T^*]$. Notice that

$$a(\sqrt{4X(t)}) \geq a(\sqrt{4X(0)}) = a(R) = M,$$

hence by (12), we have

$$\frac{d}{dt} V(t) \leq -2MV(t)$$

which, by integration, implies that $\sqrt{V(t)} \leq v^0 e^{-Mt}$ for every $t \in [0, T^*]$.

We now plug together the two bounds. In order for (15) to hold at sometime $T < T^*$, simply choose

$$M = M_T \triangleq \frac{1}{T} \log \left(\frac{2v^0}{\varepsilon} \right).$$

For this choice of M , it follows

$$\int_{\sqrt{X(T)}}^{+\infty} a(\sqrt{4r}) dr \geq \frac{\varepsilon}{2} = v^0 e^{-M_T T} \geq \sqrt{V(T)}.$$

From Theorem 2, we can then conclude that any solution of the above system tends autonomously to consensus.

4 The Effect of Perturbations on Consensus Emergence

An immediate way to enhance the alignment capabilities of systems like (7) consists in adding a feedback term penalizing the distance of each agent's velocity from the average one, i.e.,

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|) (v_j(t) - v_i(t)) + \gamma(\bar{v}(t) - v_i(t)), \end{cases} \quad (16)$$

where $\gamma > 0$ is a prescribed constant, modeling the strength of the additional alignment term.

This approach to the enforcement of consensus is a particular instance of what in the literature is known as *decentralized control strategy*, which has been thoroughly studied especially for its application in the self-organization of *unmanned aerial vehicles* (UAVs) [43], congestion control in communication networks [63], and distributed sensor networks [26]. We also refer to [74] for the stability analysis of a decentralized coordination method for dynamical systems with switching underlying communication network.

As system (16) can be rewritten as (7) with the interaction kernel $a(\cdot) + \gamma$ replacing $a(\cdot)$, by Theorem 2 and Remark 3, each solution of (16) tends to consensus.

However, the apparently innocent fix of adding the extra term above has actually a huge impact on the interpretation of the model: As pointed out in [16], this approach requires that each agent must possess at every instant a *perfect information* of the whole system, since it has to correctly compute the mean velocity of the group \bar{v} in order to compute its trajectory. This condition is seldom met in real-life situations, where it is usually only possible to ask that each agent computes an approximated mean velocity vector \bar{v}_i , instead of the true \bar{v} . These considerations lead us to the model

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|) (v_j(t) - v_i(t)) + \gamma(\bar{v}_i(t) - v_i(t)). \end{cases} \quad (17)$$

In studying under which conditions the solutions of system (17) tend to consensus, it is often desirable to express the approximated feedback as a combination of a term consisting on a *true information feedback*, i.e., a feedback based on the real average \bar{v} , and a perturbation term, which models the deviation of \bar{v}_i from \bar{v} . To this end, we rewrite system (17) in the following form:

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|) (v_j(t) - v_i(t)) + \alpha(t)(\bar{v}(t) - v_i(t)) + \beta(t)\Delta_i(t), \end{cases} \quad (18)$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are two nonnegative, piecewise continuous functions, and $\Delta_i(\cdot)$ is the deviation acting on the estimate of \bar{v} by agent i (which can, of course, depend on $(x_1(t), \dots, x_N(t), v_1(t), \dots, v_N(t))$). Therefore, solutions in this context have to be understood in terms of weak solutions in the Carathéodory sense, see [44].

Remark 4 In what follows, we will not be interested in the well-posedness of system (18), but rather in finding assumptions on the functions a , α , β , and Δ_i for which we can guarantee its asymptotic convergence to consensus.

System (18) provides the advantage of encompassing all the previously introduced models, as can be readily seen:

- If $\alpha = \beta \equiv \gamma$ and $\Delta_i = v_i - \bar{v}$, or $\alpha = \beta \equiv 0$, then we recover system (7),
- The choices $\alpha \equiv \gamma$, $\Delta_i \equiv 0$ (or equivalently $\beta \equiv 0$) yield system (16), and
- If $\alpha = \beta \equiv \gamma$ and $\Delta_i = \bar{v}_i - \bar{v}$, we obtain system (17).

The introduction of the perturbation term in system (18) may deeply modify the nature of the original model: For instance, an immediate consequence is that the mean velocity of the system is, in general, no longer a conserved quantity.

Proposition 2 For system (18), with perturbations given by the vector-valued function $\Delta(\cdot) = (\Delta_1(\cdot), \dots, \Delta_N(\cdot))$, for every $t \geq 0$ it holds

$$\frac{d}{dt}\bar{v}(t) = \beta(t)\bar{\Delta}(t).$$

Remark 5 As we have already pointed out, it is possible to recover system (7) by setting $\Delta_i = v_i - \bar{v}$, whereas we can recover system (16) for the choice $\Delta_i \equiv 0$. Note that in both cases, we have $\bar{\Delta}(t) = 0$ for every $t \geq 0$; therefore, the mean velocity is conserved both in systems (7) and (16).

We also highlight the fact that \bar{v} is not conserved even in the case that for every $t \geq 0$, and for every $i = 1, \dots, N$, we have $\Delta_i(t) = w$, where $w \in \mathbb{R}^d \setminus \{0\}$, i.e., the case in which all agents make the same mistake in evaluating the mean velocity.

4.1 General Results for Consensus Stabilization Under Perturbations

The following is a generalization of Lemma 1 to systems like (18).

Lemma 2 ([11, Lemma 3.1]) *Let $(x(\cdot), v(\cdot))$ be a solution of system (18). For every $t \geq 0$ it holds*

$$\frac{d}{dt} V(t) \leq -2a \left(\sqrt{2N X(t)} \right) V(t) - 2\alpha(t)V(t) + \frac{2\beta(t)}{N} \sum_{i=1}^N \Delta_i(t) \cdot v_i^\perp(t). \quad (19)$$

Proof Differentiating V for every $t \geq 0$, we have

$$\frac{d}{dt} V(t) = \frac{2}{N} \sum_{i=1}^N \frac{d}{dt} v_i^\perp(t) \cdot v_i^\perp(t) = \frac{2}{N} \sum_{i=1}^N \frac{d}{dt} v_i(t) \cdot v_i^\perp(t) - \frac{2}{N} \sum_{i=1}^N \frac{d}{dt} \bar{v}(t) \cdot v_i^\perp(t).$$

Hence, inserting the expression for $\dot{v}_i(t)$, using the fact that a is nonincreasing, and invoking Proposition 2, we get (19).

Since we are interested in the case where Δ_i plays an active role in the dynamics, in what follows, we assume $\beta(t) > 0$ for all $t \geq 0$. As a direct consequence of Lemma 2, we get that by controlling the magnitude of the deviations Δ_i , we can establish the unconditional convergence to consensus.

Theorem 3 *Let $(x(\cdot), v(\cdot))$ be a solution of system (18), and suppose that there exists a $T \geq 0$ such that for every $t \geq T$,*

$$\sum_{i=1}^N \Delta_i(t) \cdot v_i^\perp(t) \leq \phi(t) \sum_{i=1}^N \|v_i^\perp(t)\|^2 \quad (20)$$

for some function $\phi : [T, +\infty) \rightarrow [0, \ell]$, where

$$\ell < \frac{\min_{t \geq T} \alpha(t)}{\max_{t \geq T} \beta(t)}. \quad (21)$$

Then $(x(\cdot), v(\cdot))$ tends to consensus.

Proof Under the assumption (20), for every $t \geq T$, the upper bound in (19) can be simplified to

$$\frac{d}{dt} V(t) \leq 2\beta(t) \left(\ell - \frac{\alpha(t)}{\beta(t)} \right) V(t).$$

Integrating between T and t (where $t \geq T$), we get $V(t) \leq V(T) e^{2 \int_T^t \beta(s) \left(\ell - \frac{\alpha(s)}{\beta(s)} \right) ds}$, and as the factor $\ell - \alpha(s)/\beta(s)$ is negative while β is nonnegative, V approaches 0 exponentially fast.

We then immediately get the following

Corollary 2 *If there exists $T \geq 0$ such that $\Delta_i^\perp(t) = 0$ for every $t \geq T$ and for every $1 \leq i \leq N$, then any solution of system (18) tends to consensus.*

Proof Noting that $\Delta_i^\perp = 0$ implies $\Delta_i = \bar{\Delta}$, by (11), we have $\sum_{i=1}^N \Delta_i(t) \cdot v_i^\perp(t) = \sum_{i=1}^N \bar{\Delta} \cdot v_i^\perp(t) = 0$. Hence, we can apply Theorem 3 with $\phi(t) = 0$ for every $t \geq T$ to obtain the result.

Remark 6 A trivial implication of Corollary 2 is that any solution of system (16) tends to consensus (this was already a consequence of Theorem 2), but has moreover a rather nontrivial implication: Also, any solution of systems subjected to *deviated uniform control*, i.e., systems like (18) where $\Delta_i(t) = \Delta(t)$ for every $i = 1, \dots, N$ and for every $t \geq 0$, tends to consensus, because it holds

$$\Delta_i^\perp(t) = \Delta_i(t) - \frac{1}{N} \sum_{j=1}^N \Delta_j(t) = \Delta(t) - \Delta(t) = 0$$

for every $i = 1, \dots, N$ and for every $t \geq 0$, therefore, Corollary 2 applies. This means that systems of this kind converge to consensus even if the agents have an incorrect knowledge of the mean velocity, provided they all make the same mistake.

Another consequence of the previous results is the following corollary, which provides an upper bound for tolerable perturbations under which consensus emergence can be unconditionally guaranteed.

Corollary 3 *For every $i = 1, \dots, N$, let $\varepsilon_i : \mathbb{R}_+ \rightarrow [0, \ell]$ for $\ell > 0$ as in (21). If there exists $T \geq 0$ such that $\|\Delta_i(t)\| \leq \varepsilon_i(t) \|v_i^\perp(t)\|$ for every $t \geq T$ and for every $i = 1, \dots, N$, then any solution of system (18) tends to consensus.*

4.2 Perturbations as Leader-Based Feedback

We now consider the problem of consensus stabilization based on a leader-following feedback.

Example 3 Let us use Lemma 2 to study the convergence to consensus of a system like (17), where each agent computes its local mean velocity \bar{v}_i by taking into account

itself plus a single common agent (x_1, v_1) , which in turn takes into account only itself by computing $\bar{v}_1 = v_1$. Formally, given two finite conjugate exponents p, q (i.e., two positive real numbers satisfying $1/p + 1/q = 1$), we assume that for any $i = 1, \dots, N$, it holds

$$\bar{v}_i(t) = \frac{1}{p}v_i(t) + \frac{1}{q}v_1(t) \quad \text{for every } t \geq 0.$$

We shall prove that any solution of this system tends to consensus, no matter how small the positive weight $1/q$ of v_1 in \bar{v}_i is. We start by writing the system under the form (18), with $\alpha(t) = \beta(t) = \gamma > 0$ and

$$\Delta_i(t) = \frac{1}{p}v_i^\perp(t) + \frac{1}{q}v_1^\perp(t) \quad \text{for every } t \geq 0.$$

Hence, the perturbation term in the estimate (19) on the decay of V becomes

$$\begin{aligned} \frac{2\gamma}{N} \sum_{i=1}^N \Delta_i(t) \cdot v_i^\perp(t) &= \frac{2\gamma}{N} \sum_{i=1}^N \frac{1}{p}(v_i^\perp(t) + \frac{1}{q}v_1^\perp(t)) \cdot v_i^\perp(t) \\ &= \frac{1}{p} \frac{2\gamma}{N} \sum_{i=1}^N \|v_i^\perp(t)\|^2 + \frac{1}{q} \frac{2\gamma}{N} v_1^\perp(t) \cdot \underbrace{\sum_{i=1}^N v_i^\perp(t)}_{=0} = \frac{2\gamma}{p} V(t). \end{aligned}$$

Lemma 2 let us bound the growth of V as

$$\frac{d}{dt}V(t) \leq 2\gamma \left(-1 + \frac{1}{p} \right) V(t) = -\frac{2\gamma}{q} V(t),$$

which ensures the exponential decay of the functional V for any $q > 0$.

Figure 3 shows the behavior of the group of agents considered in Example 3 depending on the parameter $q \in (0, 1]$, which represents the influence of the leader in the local average. The result above asserts that for every such q , the system will converge to consensus independently of the initial configuration, as illustrated in Figures 3 and 4. It can be observed that the weaker the influence of the leader, the longer the group of agents takes to align.

Besides the Cucker–Smale model, the leader-following control problem was also studied in [81] for the Hegselmann–Krause model and in [12] for the D’Orsogna et al. model (see [41] as a reference): In these papers, the leader’s optimal strategy to induce pattern formation was discussed.

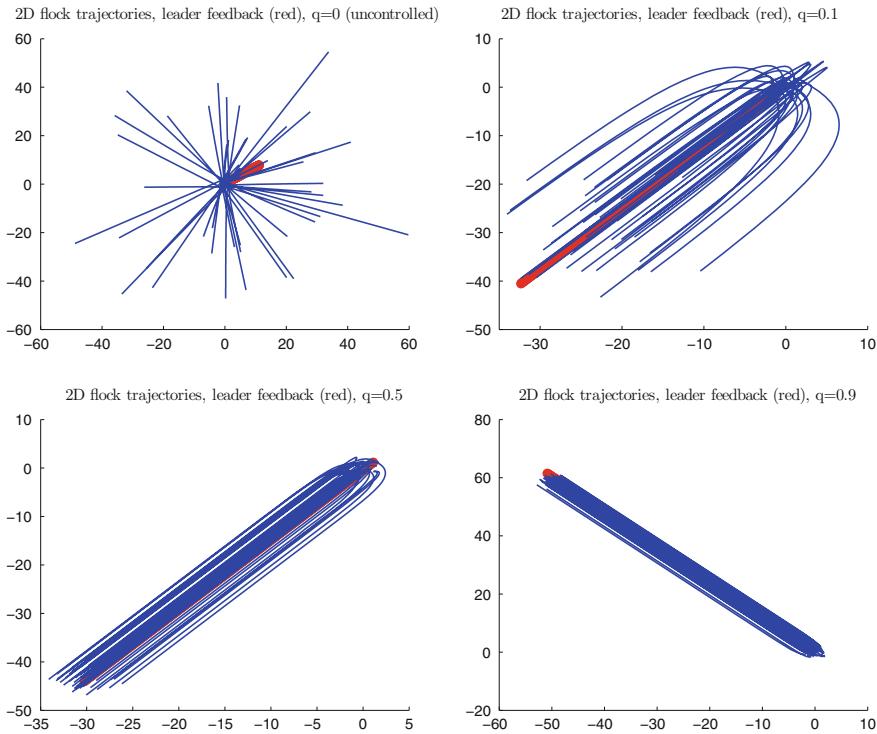


Fig. 3 Leader-based feedback control. Simulations with 100 agents, where the value q indicates the strength of the leader in the partial average. It can be observed how, as the strength of the leader is increased, convergent behavior is improved.

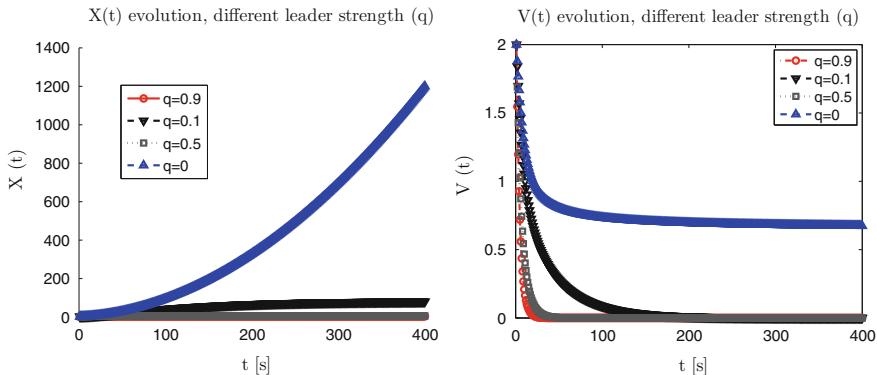


Fig. 4 Leader-based feedback control. Simulations with 100 agents, where q indicates the strength of the leader in the partial average. Evolution of X and V for the simulations in Figure 3.

4.3 Feedback Under Perturbed Information

Motivated by the example of the last section, we turn our attention to the study of systems like (18) where the perturbation of the mean of the i th agent has the specific form

$$\Delta_i(t) = \sum_{j=1}^N \omega_{ij}(t) v_j^\perp(t) \quad \text{for every } t \geq 0, \quad (22)$$

for some positive measurable mapping $\omega : \mathbb{R}_+ \rightarrow \mathbb{R}^{N \times N}$, i.e., for every $t \geq 0$, the function ω has the property $\omega_{ij}(t) > 0$ for all $i, j = 1, \dots, N$.

An example of the above framework is provided by a weight matrix of the form

$$\omega_{ij}(t) \triangleq \frac{\phi(\|x_i(t) - x_j(t)\|)}{\eta_i(t)} \quad \text{for every } t \geq 0.$$

where the weighting function ϕ corresponds to the Cucker–Smale kernel (8) with $H = 1$, $\sigma = 1$, and $\beta = \varepsilon$, i.e.,

$$\phi(r) \triangleq \frac{1}{(1 + r^2)^\varepsilon}.$$

and the normalizing terms η_i are defined as

$$\eta_i(t) \triangleq \sum_{j=1}^N \phi(\|x_i(t) - x_j(t)\|). \quad (23)$$

Let us consider the case $\alpha(t) = \alpha > 0$ and $\beta(t) = \beta > 0$ for every $t \geq 0$. Figures 5 and 6 show the behavior of the system when changing the balance between the constants α and β . In this test, we fix a large value of $\beta = 10$, representing a strong perturbation of the feedback, and a small value of $\varepsilon = 1e-5$, related to a disturbance which is distributed among all the agents: Increasing the value of α in system (18) (which represents the energy of the *correct information feedback*) induces faster consensus emergence.

As already mentioned in Section 3, the use of a common normalizing factor η in place of different terms η_i greatly helps in the study of consensus emergence. For this particular case, we get the following result.

Corollary 4 ([11, Corollary 3]) *Suppose that for all $i, j = 1, \dots, N$ the function $\omega_{ij} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ satisfies*

$$\omega_{ij}(t) = \frac{\phi(\|x_i(t) - x_j(t)\|)}{\eta(t)} \quad \text{for every } t \geq 0,$$

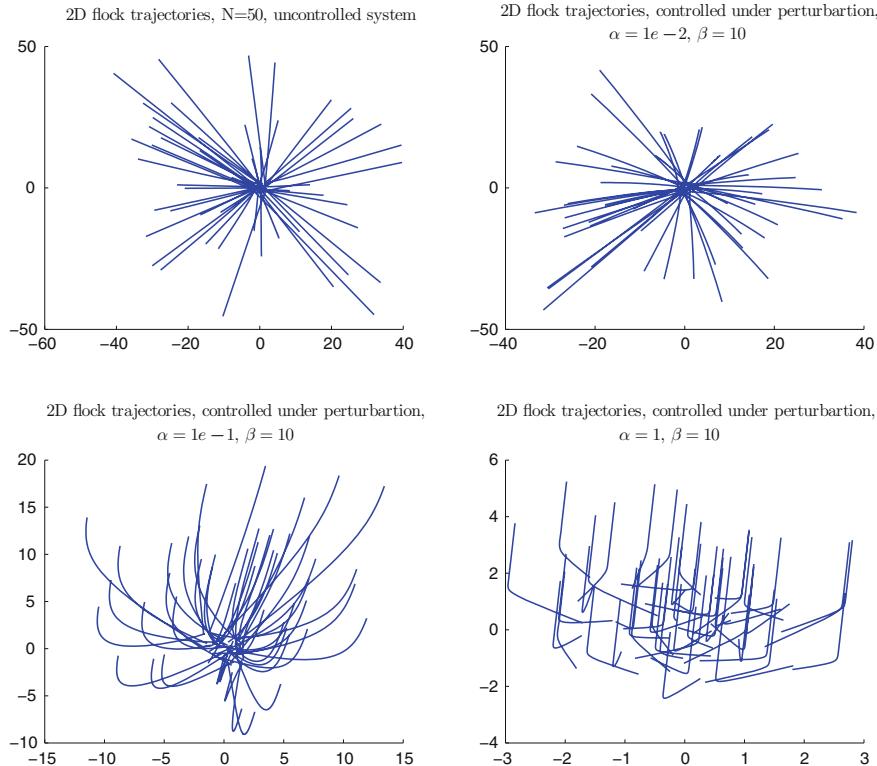


Fig. 5 Total feedback control under structured perturbations. For a fixed strong structured perturbation term ($\beta = 10$), different energies for the unperturbed control term α generate different consensus behavior; the stronger the correct information term is, the faster the consensus is achieved.

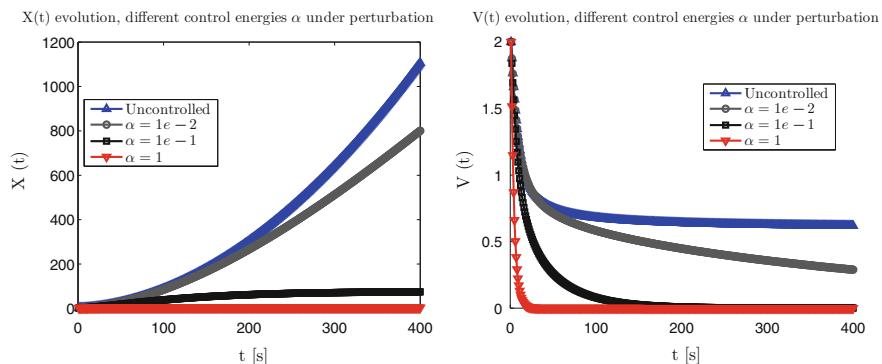


Fig. 6 Total feedback control under structured perturbations. Evolution of X and V for the simulations in Figure 5.

where $\phi : \mathbb{R}_+ \rightarrow (0, 1]$ is a nonincreasing, positive, bounded function, and $\eta : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a nonnegative bounded function satisfying

$$1 \leq N \frac{\beta(t)}{\alpha(t)} \leq \eta(t) \quad \text{for every } t \geq 0.$$

Then, any solution of system (18) with Δ_i as in (22) tends to consensus.

Remark 7 A concrete example of a system for which we can apply Corollary 4 is obtained by considering the common normalizing term

$$\eta(t) \stackrel{\Delta}{=} \max_{1 \leq i \leq N} \left\{ \sum_{j=1}^N \phi(\|x_i(t) - x_j(t)\|) \right\}.$$

which is also coherent with the asymptotic behavior of (23) for $\varepsilon \rightarrow 0$ and $\varepsilon \rightarrow +\infty$.

Remark 8 The request of positivity of the function ϕ cannot be removed from Corollary 4, see [11, Remark 4]

4.4 Perturbations Due to Local Averaging

An interesting case of a system like (17) is the one where the local mean is given by

$$\bar{v}_i(t) = \frac{1}{|\Lambda_R(t, i)|} \sum_{j \in \Lambda_R(t, i)} v_j(t) \quad \text{for every } t \geq 0, \quad (24)$$

where $\Lambda_R(i)$ is defined as in (5). In this case, we model the situation in which each agent estimates the average velocity of the group in the extra feedback term by only counting those agents inside a ball of radius R centered on him.

Simulations in Figure 7 illustrate the behavior of such configuration. From an uncontrolled system, represented by a local feedback radius $R = 0$, by increasing this quantity, partial alignment is consistently achieved, until full consensus is observed for large radii mimicking a total information feedback control.

We want to address the issue of characterizing the behavior of system (17) with the above choice for \bar{v}_i when the radius R of each ball is either reduced to 0 or set to grow to $+\infty$. We shall see that we can reformulate this decentralized system again as a Cucker–Smale model for a different interaction function for which we can apply Theorem 2. We shall show how tuning the radius R affects the convergence to consensus, from the case $R \geq 0$, where only conditional convergence is ensured, to the unconditional convergence result given for $R = +\infty$.

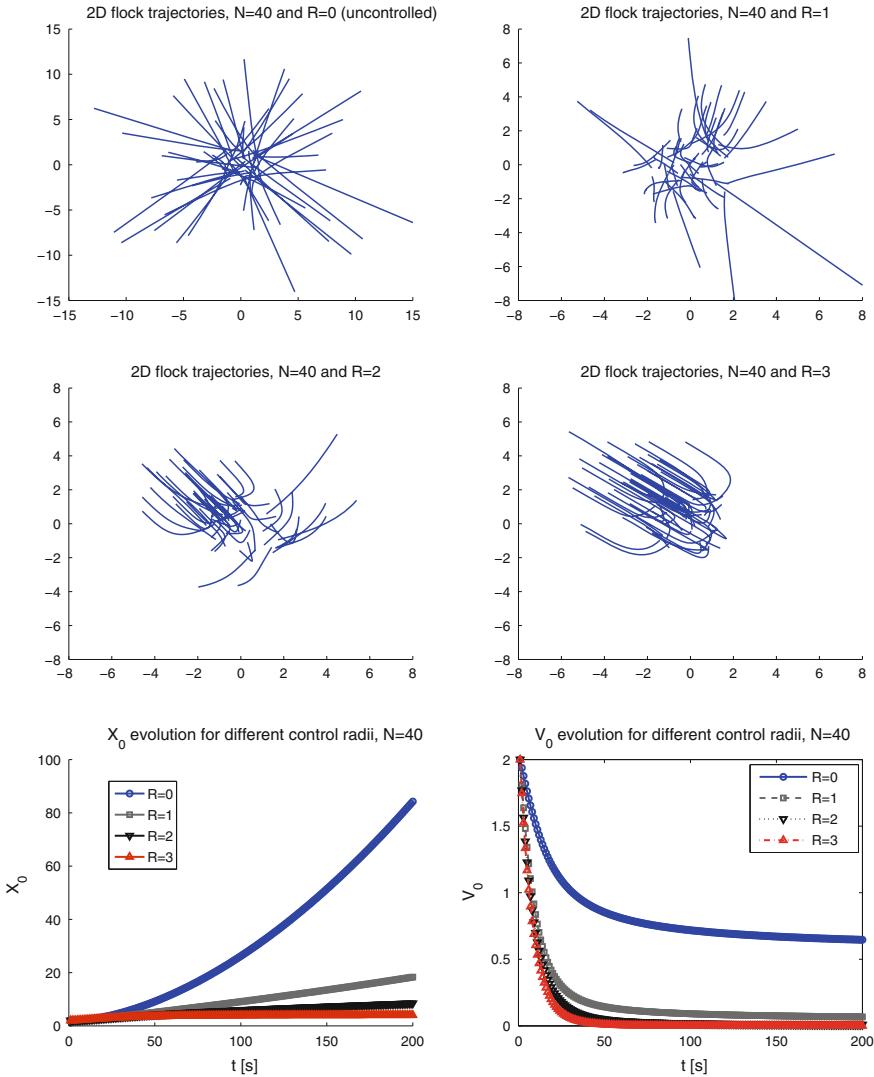


Fig. 7 Local feedback control. Simulations with $N = 40$ agents, and different control radii R . By increasing the value of R , the systems transit from uncontrolled behavior to partial alignment, up to total, fast alignment.

4.4.1 Preserving the Asymptotics

First of all, by means of $\chi_{[0, R]}$, we can rewrite $\bar{v}_i(t)$ as

$$\bar{v}_i(t) = \frac{1}{\sum_{k=1}^N \chi_{[0, R]}(\|x_i(t) - x_k(t)\|)} \sum_{j=1}^N \chi_{[0, R]}(\|x_i(t) - x_j(t)\|) v_j(t). \quad (25)$$

Unfortunately, the normalizing terms $\sum_{k=1}^N \chi_{[0, R]}(\|x_i(t) - x_k(t)\|)$ give rise to a matrix of weights which is not symmetric, which greatly complicates the analysis of the convergence to consensus. However, since we are mainly interested in the limit behavior of the system for $R \rightarrow 0$ and $R \rightarrow +\infty$, following Remark 7 we take η_R to be a function approximating the above normalizing terms and which also preserves its asymptotics for $R \rightarrow 0$ and $R \rightarrow +\infty$, as for instance,

$$\eta_R(t) = \max_{1 \leq i \leq N} \left\{ \sum_{k=1}^N \chi_{[0, R]}(\|x_i(t) - x_k(t)\|) \right\}. \quad (26)$$

Therefore, we replace the vector $\bar{v}_i(t)$ by

$$\frac{1}{\eta_R(t)} \sum_{j=1}^N \chi_{[0, R]}(\|x_i(t) - x_j(t)\|) v_j(t).$$

On top of this, notice that the vector

$$\left(\frac{1}{\eta_R(t)} \sum_{j=1}^N \chi_{[0, R]}(\|x_i(t) - x_j(t)\|) \right) v_i(t)$$

is an approximation of $v_i(t)$ for $R \rightarrow 0$ and $R \rightarrow +\infty$. This motivates the replacement of the term $\bar{v}_i - v_i$ where \bar{v}_i is as in (25) with

$$\bar{v}_i - v_i \approx \frac{1}{\eta_R} \sum_{j=1}^N \chi_{[0, R]}(\|x_i - x_j\|) v_j - \left(\frac{1}{\eta_R} \sum_{j=1}^N \chi_{[0, R]}(\|x_i - x_j\|) \right) v_i. \quad (27)$$

The term (27) can be rewritten as $1/\eta_R \sum_{j=1}^N \chi_{[0, R]}(\|x_i - x_j\|)(v_j - v_i)$, which can be further simplified as follows

$$\begin{aligned}
\frac{1}{\eta_R} \sum_{j=1}^N \chi_{[0,R]}(r_{ij})(v_j - v_i) &= \frac{1}{\eta_R} \sum_{i=1}^N (v_j - v_i) + \frac{1}{\eta_R} \sum_{j=1}^N (1 - \chi_{[0,R]}(r_{ij}))(v_i - v_j) \\
&= \frac{N}{\eta_R} (\bar{v} - v_i) + \frac{1}{\eta_R} \sum_{j=1}^N (1 - \chi_{[0,R]}(r_{ij}))(v_i - v_j),
\end{aligned} \tag{28}$$

where we have written r_{ij} in place of $\|x_i - x_j\|$ and removed the time dependencies for the sake of compactness.

It is clear that the choice of the function $\chi_{[0,R]}$ is arbitrary and other alternatives can be selected, provided they give a coherent approximation of the local average (24). For instance, instead of $\chi_{[0,R]}$ and η_R , we can consider two generic functions ψ_ε and η_ε , where ε is a parameter ranging in a nonempty set Ω , satisfying the following properties:

- (i) $\psi_\varepsilon : \mathbb{R}_+ \rightarrow [0, 1]$ is a nonincreasing measurable function for every $\varepsilon \in \Omega$;
- (ii) $\eta_\varepsilon \in L^\infty(\mathbb{R}_+)$ for every $\varepsilon \in \Omega$; and
- (iii) There are two disjoint subsets Ω_{CS} and Ω_U of Ω such that
 - if $\varepsilon \in \Omega_{CS}$, then $\psi_\varepsilon = \chi_{\{0\}}$ and $\eta_\varepsilon \equiv 1$ and
 - if $\varepsilon \in \Omega_U$, then $\psi_\varepsilon = \chi_{\mathbb{R}_+}$ and $\eta_\varepsilon \equiv N$.

Under the above hypotheses, we consider the perturbation given for every $t \geq 0$ by

$$\Delta_i^\varepsilon(t) \triangleq \frac{1}{\eta_\varepsilon(t)} \sum_{j=1}^N (1 - \psi_\varepsilon(\|x_i(t) - x_j(t)\|))(v_i(t) - v_j(t)). \tag{29}$$

With requirement (iii), we impose that whenever $\varepsilon \in \Omega_{CS}$, then it holds

$$\Delta_i^\varepsilon(t) = -\frac{N}{\eta_\varepsilon(t)} (\bar{v}(t) - v_i(t)),$$

therefore recovering the Cucker–Smale system (7) from (30), while whenever $\varepsilon \in \Omega_U$, then $\Delta_i^\varepsilon(t) = 0$ holds, and we obtain a particular instance of system (16).

4.4.2 The Enlarged Consensus Region

By means of (28) and (29), we can rewrite our system with the local average (24) in the form of system (18) as follows

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|) (v_j(t) - v_i(t)) + \gamma \frac{N}{\eta_\varepsilon(t)} (\bar{v}(t) - v_i(t)) + \gamma \Delta_i^\varepsilon(t). \end{cases} \quad (30)$$

Using (28) and collecting the term $(v_j - v_i)$, it is easy to see that Theorem 2 yields the following description of the consensus region as a function of the parameter ε .

Theorem 4 Fix $\gamma \geq 0$, consider system (30) where Δ_i^ε is as in (29) and let $(x^0, v^0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$. If $X_0 \stackrel{\Delta}{=} B(x^0, x^0)$ and $V_0 \stackrel{\Delta}{=} B(v^0, v^0)$ satisfy

$$\int_{\sqrt{X_0}}^{+\infty} a(\sqrt{2Nr}) dr + \frac{\gamma N}{\|\eta_\varepsilon\|_{L^\infty(\mathbb{R}_+)}} \int_{\sqrt{X_0}}^{+\infty} \psi_\varepsilon(\sqrt{2Nr}) dr \geq \sqrt{V_0}, \quad (31)$$

then the solution of system (30) with initial datum (x^0, v^0) tends to consensus.

Let us see how we can apply Theorem 4 to obtain an estimate of the consensus region for the local average (24). We consider $\Omega = [0, +\infty]$, the sequence of functions $(\chi_{[0, R]})_{R \in \Omega}$ and η_R as in (26) (notice that, as before, we have $\Omega_{CS} = \{0\}$ and $\Omega_U = \{\infty\}$). Since it holds $\|\eta_R\|_{L^\infty(\mathbb{R}_+)} \leq N$, if R is sufficiently large to satisfy $\sqrt{2NX_0} \leq R$, condition (31) is satisfied as soon as

$$\int_{\sqrt{X_0}}^{+\infty} a(\sqrt{2Nr}) dr + \gamma \left(\frac{R}{\sqrt{2N}} - \sqrt{X_0} \right) \geq \sqrt{V_0},$$

by means of a trivial integration. If, instead, R is so small that $\sqrt{2NX_0} > R$ holds, condition (31) is satisfied as soon as

$$\int_{\sqrt{X_0}}^{+\infty} a(\sqrt{2Nr}) dr \geq \sqrt{V_0},$$

recovering Theorem 2. As can be seen, we have enlarged the original consensus region provided by Theorem 2 by a term whose size is linearly increasing in R . This implies that in the case $R = +\infty$, the consensus region coincides with the entire space $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$; hence, the system converges to consensus regardless of the initial datum.

4.4.3 Empirical Estimation of the Enlarged Consensus Region

We present a series of numerical tests aiming at estimating empirically the enlarged consensus region given by (31) following similar ideas as those presented in [19]. We consider a system of N agents in dimension $d = 2$ with a randomly generated initial configuration of positions and velocities

$$(x^0, v^0) \in [-1, 1]^{2N} \times [-1, 1]^{2N},$$

interacting by means of the kernel (8) with $H = 1$, $\sigma = 1$, and $\beta = 1$. We recall that relevant quantities for the analysis of our results are given by (here, we stress the dependance on x and v)

$$X[x](t) \triangleq \frac{1}{2N^2} \sum_{i,j=1}^N \|x_i(t) - x_j(t)\|^2 \quad \text{and} \quad V[v](t) \triangleq \frac{1}{2N^2} \sum_{i,j=1}^N \|v_i(t) - v_j(t)\|^2.$$

Notice that once a random initial configuration has been generated, it is possible to rescale it to a desired (X_0, V_0) parametric pair, by means of

$$(x, v) = \left(\sqrt{\frac{X_0}{X[\tilde{x}]}} \tilde{x}, \sqrt{\frac{V_0}{V[\tilde{v}]}} \tilde{v} \right),$$

such that $(X[x], V[v]) = (X_0, V_0)$. As simulations of the trajectories have been generated by prescribing a value for the pair (X_0, V_0) , which is used to rescale randomly generated initial conditions, there are slight variations on the initial positions and velocities in every model run, which can affect the final consensus direction. However, our results are stated in terms of X , V and independently of the specific initial configuration. For simulation purposes, the system is integrated in time with the specific feedback control by means of a Runge–Kutta fourth-order scheme.

As it was shown in Example 1, that estimates for consensus regions such as the one provided by Theorem 2 are not sharp in many situations. In this direction, we proceed to contrast the theoretical consensus estimates with the numerical evidence. For this purpose, for a fixed number of agents, we span a large set of possible initial configurations determined by different values of (X_0, V_0) . For every pair (X_0, V_0) , we randomly generate a set of 20 initial conditions, and we simulate for a sufficiently large time frame. We measure consensus according to a threshold established on the final value of V ; we consider that consensus has been achieved if the final value of V is lower or equal to $1e - 5$. We proceed by computing empirical probabilities of consensus for every point of our state space (X_0, V_0) ; results in this direction are shown in Figures 8 and 9. We first consider the simplified case of 2 agents; according to Example 1, for this particular case, the consensus region estimate provided by Theorem 2 is sharp, as illustrated by the results shown in Figure 8. Furthermore, it is also the case for Theorem 4; for $R > 0$, the predicted consensus region coincides with the numerically observed ones.

Figure 9 shows the case when a larger number of agents are considered. In a similar way as for Theorem 2, the consensus region estimate is conservative if compared with the region where numerical experiments exhibit convergent behavior. Nevertheless, Theorem 4 is consistent in the sense that the theoretical consensus region increases gradually as R grows, eventually covering any initial configuration, which is the case of the total information feedback control, as presented in [16, Proposition 2]. The

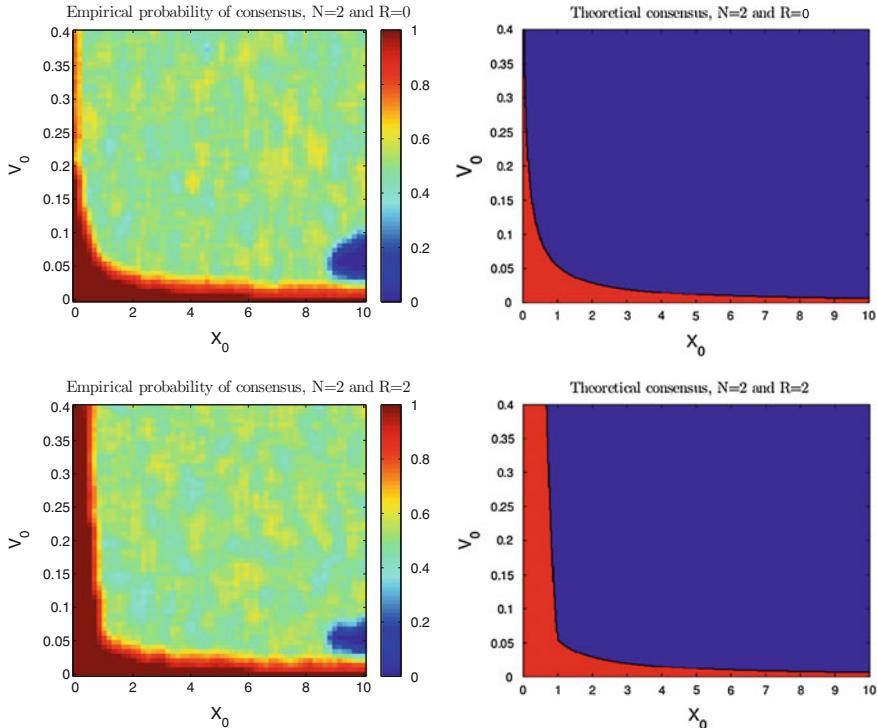


Fig. 8 Local feedback control. Empirical consensus regions and theoretical estimates for two-agent systems.

numerical experiments also confirm this phenomena, as shown in Figure 10, where contour lines showing the 80% probability of consensus for different radii locate farther from the origin as R increases.

5 Sparse Control of the Cucker–Smale Model

We have seen throughout the previous sections how difficult it is to ensure unconditional convergence to consensus for alignment models. In particular, in Section 4.4, we have proven that the addition of a local feedback does not always help: Theorem 4 shows that we can guarantee unconditional convergence to consensus with respect to the initial datum for dynamical systems of the form

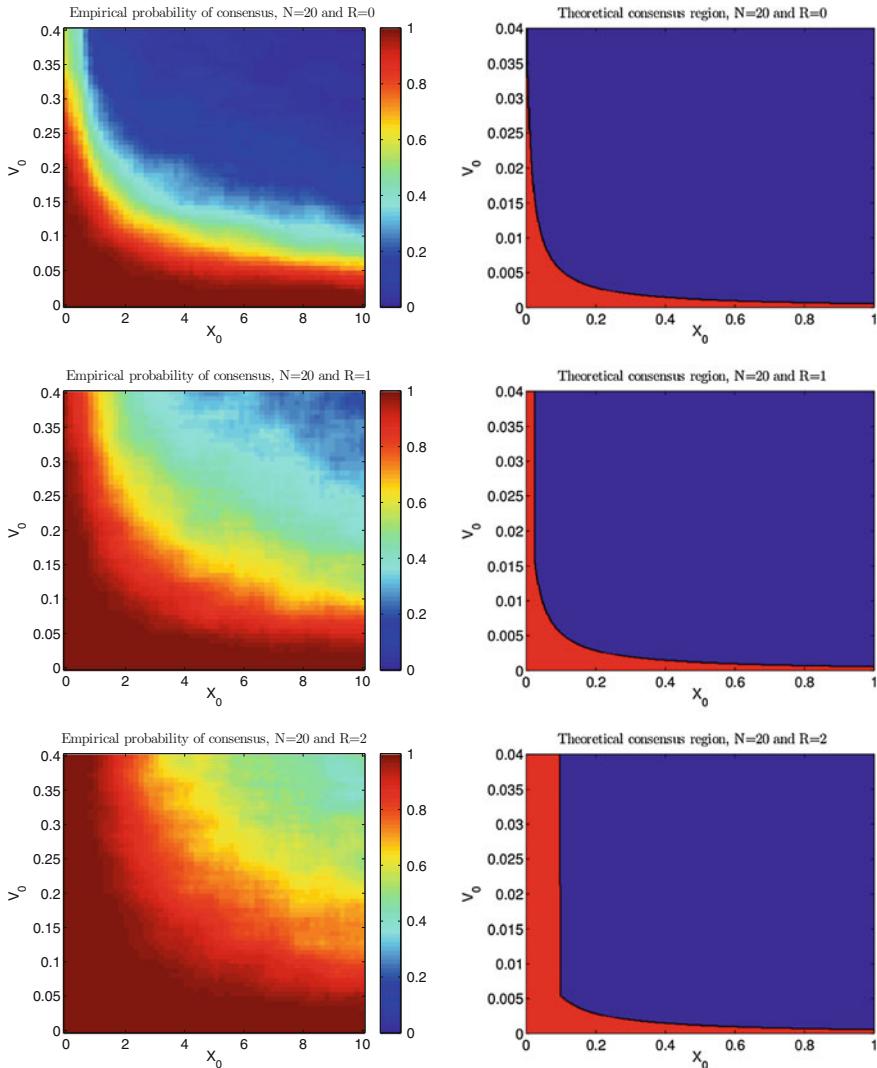
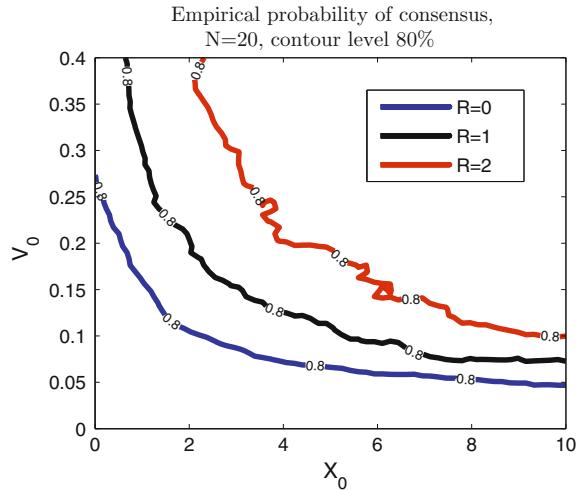


Fig. 9 Local feedback control. Empirical consensus regions and theoretical estimates for $N = 20$ agents and different control radii R .

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|) (v_j(t) - v_i(t)) + \gamma \left(\frac{1}{|A_R(t, i)|} \sum_{j \in A_R(t, i)} v_j(t) - v_i(t) \right). \end{cases}$$

only in the case $R = +\infty$, for which the identity

Fig. 10 Local feedback control. Empirical contour lines for the 80% probability of consensus with different control radii.



$$\frac{1}{|\Lambda_R(t, i)|} \sum_{j \in \Lambda_R(t, i)} v_j(t) = \bar{v}(t)$$

holds. This means that either the agents have perfect information of the state of the entire system (so that the local mean \bar{v}_i is equal to the true mean \bar{v}), or as the numerical simulations in Section 4.4.3 show, there are situations where the agents are not able to converge to consensus. As already pointed out in Section 4, this is a very strong requirement to ask for, and not many real-life scenarios are able to support it. Consider, for instance, the case of an assembly of people trying to reach an unanimous decision, like the European Union Council: Since the extra term can be interpreted as an additional desire of each agent to agree with people whose goal is near to his, the requirement $R = +\infty$ corresponds to asking that all the individual goals are close; i.e., all agents pursue the same end. A truly imaginative world indeed! We are thus facing an inherent, severe limitation of the decentralized approach.

5.1 Centralized Feedback Interventions

To overcome this apparent dead end, let us write $u_i(t) = \gamma(\bar{v}(t) - v_i(t))$, i.e.,

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = \frac{1}{N} \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|)(v_j(t) - v_i(t)) + u_i(t). \end{cases} \quad (32)$$

Instead of interpreting u_i as a *decentralized* force, let us consider it as an *external* force from an outside source acting on the system to help it to coordinate. This new approach sheds a completely different light on the problem: With respect to the example considered before, it is like introducing a moderator heading the discussion, who can make pressure on the participants to the council facilitating the consensus process. Adding an external figure implementing intervention policies broadens further the expressive power of the problem: Indeed, since we are in principle no more tied to specific interventions of the form $u_i(t) = \gamma(\bar{v}(t) - v_i(t))$, this setting enables us to ask ourselves the following question

(Q) *given a set of constraints, which control u is the best to reach a specific goal?*

In this section, we shall study a specific instance of this very general issue in the case of system (32). In our setting, the constraints shall be

- (i) The control is of *feedback type*, i.e., computed instantaneously as a function of the state variables, following a *locally optimal* criterion;
- (ii) There is a maximal amount of resources $M > 0$ that the central policy maker can spend at any given time for the intervention; and
- (iii) The control should act on the least amount of agents possible at any time.

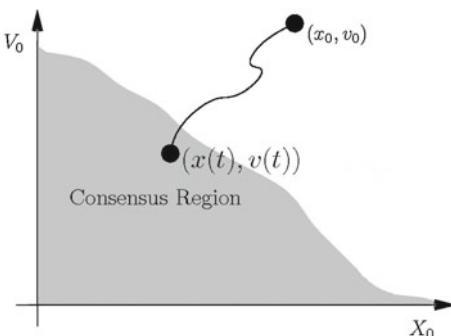
For the time being, our goal is again alignment; hence, we seek for a control u for which the associated solution with system (32) tends to consensus in the sense of Definition 3. We have seen in Proposition 1 that an effective criterion for consensus emergence is the minimization of the Lyapunov functional V : If we are able to prove that our control strategy is able to drive V below the threshold level given by Theorem 2, we have automatically consensus emergence (see Figure 11). The maximization of the decay rate of V is a locally optimal criterion and hence compatible with point (i).

The following preliminary estimate shows the effect of a control on V .

Lemma 3 *For any measurable function $u : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN}$, it holds*

$$\frac{d}{dt} V(t) \leq 2B(u(t), v(t)).$$

Fig. 11 Steering the system to a point fulfilling the conditions of Theorem 2.



Proof Using the representation of system (7) in Laplacian form (9), and the fact that $L(x(t))$ is positive definite, we get

$$\frac{d}{dt} V(t) = \frac{d}{dt} B(v(t), v(t)) = -2B(L(x(t))v(t), v(t)) + 2B(u(t), v(t)) \leq 2B(u(t), v(t)).$$

This concludes the proof.

The constraint on the maximal amount of available resources M given by point (ii) leads to the following definition of *admissible controls*.

Definition 6 (Admissible controls) A measurable function $u = (u_1, \dots, u_N) : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN}$ is an *admissible control* if it satisfies

$$\sum_{i=1}^N \|u_i(t)\| \leq M \quad \text{for every } t \geq 0. \quad (33)$$

As an immediate corollary of Lemma 3, we can show that the problem of finding admissible controls steering the system to consensus is well posed.

Corollary 5 (Total control, [16, Proposition 2]) Fix $M > 0$, an initial condition $(x^0, v^0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$, and $0 < \alpha \leq M/(N\sqrt{V_0})$. Then, the feedback control defined pointwise in time as

$$u(t) = -\alpha v^\perp(t) \quad \text{for every } t \geq 0, \quad (34)$$

is admissible and the solution associated to u tends to consensus.

Proof Let $(x, v) : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ be a solution of system (32) with u as in the statement. Lemma 3 implies that

$$\frac{d}{dt} V(t) \leq 2B(u(t), v(t)) = -2\alpha B(v^\perp(t), v(t)) = -2\alpha V(t);$$

Therefore, an application of Gronwall's Lemma yields $V(t) \leq e^{-2\alpha t} V(0)$, so $V(t)$ tends to 0 exponentially fast as $t \rightarrow +\infty$. In particular, $X(t)$ keeps bounded and the trajectory reaches the consensus region in finite time. Lastly, it follows that

$$\sum_{i=1}^N \|u_i(t)\| \leq \sqrt{N} \sqrt{\sum_{i=1}^N \|u_i(t)\|^2} = \alpha \sqrt{N} \sqrt{\sum_{i=1}^N \|v_i^\perp(t)\|^2} = \alpha N \sqrt{V(t)} \leq \alpha N \sqrt{V_0} \leq M,$$

which implies the admissibility of the control.

Corollary 5, although very simple, is somehow remarkable: It shows not only that we can steer to consensus the system from any initial condition, but also that the strength of the control $M > 0$ can be arbitrarily small. However, this result has

perhaps only theoretical validity, because the stabilizing control $u = -\alpha v^\perp$ needs to act instantaneously on all the agents, thus requiring the external policy maker to interact at every instant with all the agents in order to steer the system to consensus, a procedure that requires a large amount of instantaneous communications, whence the name of *total control*. This motivates point (iii) and is the reason why we look for interventions that target the fewest number of agents at any given time. However, this leads us into the difficult combinatorial problem of the selection of the best few control components to be activated. How can we solve it?

The problem resembles very much the one in information theory of finding the best possible sparse representation of data in the form of vector coefficients with respect to an adapted dictionary for the sake of their compression, see [56, Chapter 1]. In our case, the relationship between control choices and result will be usually highly non-linear, especially for several known dynamical systems modeling social dynamics: Were this relationship more simply linear instead, then a rather well-established theory would predict how many degrees of freedom are minimally necessary to achieve the expected outcome. Moreover, depending on certain spectral properties of the linear model, the theory allows also for efficient algorithms to compute the relevant degrees of freedom, relaxing the associated combinatorial problem. This theory is known in mathematical signal processing and information theory under the name of *compressed sensing*, see the seminal work [14, 40] and the review chapter [46]. The major contribution of these papers was to realize that one can combine the power of convex optimization, in particular ℓ_1 -norm minimization, and spectral properties of random linear models in order to achieve optimal results on the ability of ℓ_1 -norm minimization of recovering robustly linearly constrained sparsest solutions. Borrowing a leaf from compressed sensing, we model sparse stabilization and control strategies by penalizing the class of vector-valued controls $u = (u_1, \dots, u_N) \in \mathbb{R}^{dN}$ by means of the mixed $\ell_1^N - \ell_2^d$ -norm

$$\sum_{i=1}^N \|u_i\|_{\ell_2^d}$$

The above mixed norm has been already used, for instance, in [42] to optimally sparsify multivariate vectors in compressed sensing problems or in [45] as a *joint sparsity* constraint. The use of ℓ_1 -norms to penalize controls was first introduced in the seminal paper [29] to model linear fuel consumption, while lately the use of L^1 minimization in optimal control problems with partial differential equation has become very popular, for instance in the modeling of optimal placing of sensors [21, 24, 52, 73, 79].

5.2 Sparse Feedback Controls

We wonder whether we can stabilize the system by means of interventions that are more parsimonious than the total control, since they are more realistically modeling actual government actions. From Lemma 3, we learn that a good strategy to steer the system to consensus is actually the minimization of $B(u(t), v(t))$ with respect to u , for all t . For this reason, we choose controls according to a specific variational principle leading to a componentwise sparse stabilizing feedback law.

Definition 7 For every $M > 0$ and every $(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$, let $U(x, v)$ be the set of solutions of the variational problem

$$\min_{u \in \mathbb{R}^{dN}} \left(B(u, v) + \gamma(B(x, x)) \frac{1}{N} \sum_{i=1}^N \|u_i\| \right) \quad \text{subject to } \sum_{i=1}^N \|u_i\| \leq M, \quad (35)$$

where the *threshold functional* γ is defined as

$$\gamma(X) \triangleq \int_{\sqrt{X}}^{\infty} a(\sqrt{2Nr}) dr.$$

Notice that the variational principle (35) is balancing the minimization of $B(u, v)$, which we mentioned above as relevant to promote convergence to consensus, and the ℓ_1 -norm term $\sum_{i=1}^N \|u_i\|$, expected to promote sparsity.

Each value of $\gamma(B(x, x))$ yields a partition of $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$ into four disjoint sets:

$$\begin{aligned} \mathcal{P}_1 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i^\perp\| < \gamma(B(x, x))^2\}, \\ \mathcal{P}_2 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i^\perp\| = \gamma(B(x, x))^2 \text{ and } \exists k \geq 1 \text{ and} \\ &\quad i_1, \dots, i_k \in \{1, \dots, N\} \text{ such that } \|v_{i_1}^\perp\| = \dots = \|v_{i_k}^\perp\| \text{ and } \|v_{i_1}^\perp\| > \|v_j^\perp\| \\ &\quad \text{for every } j \notin \{i_1, \dots, i_k\}\}, \\ \mathcal{P}_3 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i^\perp\| > \gamma(B(x, x))^2 \text{ and } \exists i \in \\ &\quad \{1, \dots, N\} \text{ such that } \|v_i^\perp\| > \|v_j^\perp\| \text{ for every } j \neq i\}, \\ \mathcal{P}_4 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i^\perp\| > \gamma(B(x, x))^2 \text{ and } \exists k > 1 \text{ and} \\ &\quad i_1, \dots, i_k \in \{1, \dots, N\} \text{ such that } \|v_{i_1}^\perp\| = \dots = \|v_{i_k}^\perp\| \text{ and } \|v_{i_1}^\perp\| > \|v_j^\perp\| \\ &\quad \text{for every } j \notin \{i_1, \dots, i_k\}\}, \end{aligned}$$

Moreover, since we are minimizing $B(u, v) = B(u, v^\perp)$, it is easy to see that for every $(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ and every element $u(x, v) = (u_1(x, v), \dots, u_N(x, v))^T \in U(x, v)$, there exist nonnegative real numbers $\varepsilon_i \geq 0$ such that, for every $i = 1, \dots, N$, it holds

$$u_i(x, v) = \begin{cases} -\varepsilon_i \frac{v_i^\perp}{\|v_i^\perp\|} & \text{if } \|v_i^\perp\| \neq 0, \\ 0 & \text{if } \|v_i^\perp\| = 0, \end{cases} \quad (36)$$

where $0 \leq \sum_{i=1}^N \varepsilon_i \leq M$. The values of the ε_i 's can be determined on the basis of which partition (x, v) belongs to:

- If $(x, v) \in \mathcal{P}_1$, then $\varepsilon_i = 0$ for every $i = 1, \dots, N$;
- If $(x, v) \in \mathcal{P}_2$, then indicating with i_1, \dots, i_k the indexes such that $\|v_{i_1}^\perp\| = \dots = \|v_{i_k}^\perp\| = \gamma(B(x, x))$ and $\|v_{i_1}^\perp\| > \|v_j^\perp\|$ for every $j \notin \{i_1, \dots, i_k\}$, we have $\varepsilon_j = 0$ for every $j \notin \{i_1, \dots, i_k\}$;
- If $(x, v) \in \mathcal{P}_3$, then, indicating with i the only index such that $\|v_i^\perp\| > \|v_j^\perp\|$ for every $j \neq i$, we have $\varepsilon_i = M$ and $\varepsilon_j = 0$ for every $j \neq i$; and
- If $(x, v) \in \mathcal{P}_4$, then, indicating with i_1, \dots, i_k the indexes such that $\|v_{i_1}^\perp\| = \dots = \|v_{i_k}^\perp\|$ and $\|v_{i_1}^\perp\| > \|v_j^\perp\|$ for every $j \notin \{i_1, \dots, i_k\}$, we have $\varepsilon_j = 0$ for every $j \notin \{i_1, \dots, i_k\}$ and $\sum_{\ell=1}^k \varepsilon_{i_\ell} = M$.

Notice that any control $u(x, v) \in U(x, v)$ acts as an additional force which pulls agents toward having the same mean consensus parameter. The imposition of the $\ell_1^N - \ell_2^d$ -norm constraint has the function of enforcing *sparsity*: from the observation above clearly follows that

$$U|_{\mathcal{P}_1} = \{0\} \quad \text{and} \quad U|_{\mathcal{P}_3} = \{(0, \dots, 0, -Mv_i^\perp/\|v_i^\perp\|, 0, \dots, 0)^T\},$$

for some unique $i \in \{1, \dots, N\}$, i.e., the restrictions of U to \mathcal{P}_1 and to \mathcal{P}_3 are single valued. However, even if not all controls belonging to U are sparse, there exist selections with maximal sparsity.

Definition 8 ([16, Definition 4]) We select the *sparse feedback control* $u(x, v) \in U(x, v)$ according to the following criterion:

- if $\max_{1 \leq i \leq N} \|v_i^\perp\| \leq \gamma(B(x, x))^2$, then $u(x, v) = 0$;
- if $\max_{1 \leq i \leq N} \|v_i^\perp\| > \gamma(B(x, x))^2$, denote with $\hat{i}(x, v) \in \{1, \dots, N\}$ the smallest index such that

$$\|v_{\hat{i}(x, v)}^\perp\| = \max_{1 \leq i \leq N} \|v_i^\perp\|.$$

Then

$$u_j(x, v) \triangleq \begin{cases} -M \frac{v_{\hat{i}(x, v)}^\perp}{\|v_{\hat{i}(x, v)}^\perp\|} & \text{if } j = \hat{i}(x, v), \\ 0 & \text{otherwise.} \end{cases}$$

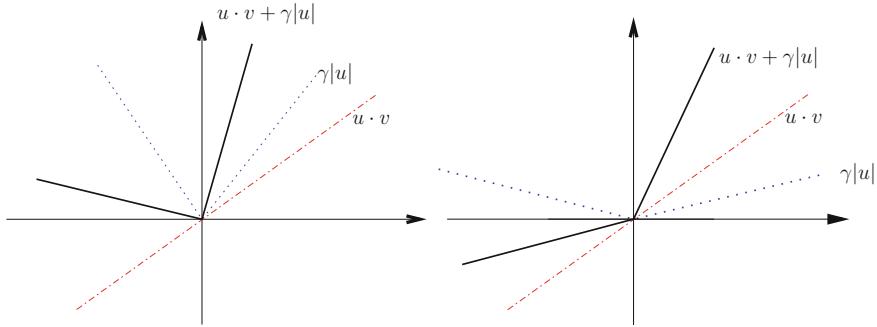


Fig. 12 Geometrical interpretation of the solution of (35) in the scalar case. On the left: for $|v| \leq \gamma$, the minimal solution $u \in [-M, M]$ is zero. On the right: for $|v| > \gamma$, the minimal solution $u \in [-M, M]$ is for $|u| = M$.

The geometrical interpretation of why the sparse feedback control is a solution of (35) is given by the graphics in Figure 12 below, representing the scalar situation.

The following result shows that the above feedback control strategy is capable of steering the system to the consensus region in finite time.

Theorem 5 ([16, Theorem 3]) *For every initial condition $(x_0, v_0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ and $M > 0$, there exist $T > 0$ and a piecewise constant selection of the sparse feedback selection of Definition 8 such that the associated absolutely continuous solution reaches the consensus region at the time T .*

This result is truly remarkable, since it holds again independently of the initial conditions and of the strength $M > 0$ of the control. Furthermore, the sparse feedback control is optimal for consensus problems with respect to any other control strategy in $U(x(t), v(t))$, which spreads control over multiple agents, as the following result shows.

Proposition 3 ([16, Proposition 3]) *The sparse feedback control of Definition 8 is for every $t \geq 0$ an instantaneous minimizer of*

$$\mathcal{D}(t, u) \triangleq \frac{d}{dt} V(t)$$

over all possible feedback controls in $U(x(t), v(t))$.

A direct consequence of Proposition 3 is that for Cucker–Smale systems, a feedback stabilization is most effective if all the attention of the controller is focused on the agent farthest away from consensus. This also means that despite the fact that the external policy maker may have few resources at disposal and can allocate them at each time only on very few key players in the system, it is always possible to effectively stabilize the dynamics to return to energy levels where the system tends autonomously to consensus. This result is perhaps surprising if confronted with the

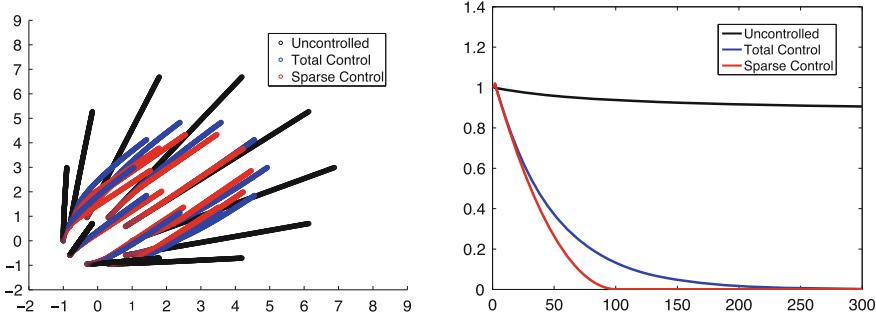


Fig. 13 Comparison between sparse, total, and no control. On the left: space evolution without control (in black), with the total control (in blue), and with the sparse control strategy (in red). On the right: the respective behavior of the functional V .

more intuitive strategy of controlling more, or even all, agents at the same time. This let us answer the question (Q) raised at the beginning of this section as follows:

(A) *under the constraints (i)–(iii), sparse is better.*

5.3 Numerical Implementation of the Sparse Control Strategy

We now compare the performances of the sparse feedback control with the self-organizing power of an uncontrolled Cucker–Smale system and the efficacy of the total control strategy (34). In Figure 13–left, it is shown a simulation of a Cucker–Smale system with $\beta = 1$ without control (in black), with the total control (in blue), and with the sparse feedback control (in red). While the uncontrolled scenario seems far from converging toward a consensus state, both the total control and the sparse control strategies successfully align the agents in very short time. The greater effectiveness of the sparse feedback control can be witnessed in Figure 13–right, where it is shown the decay of the Lyapunov functional V in the three different cases: The sparse control is more efficient in bringing V to 0, as Proposition 3 predicts.

The situation where the sparse control strategy works at its best is when the velocities of the agents are almost homogeneous, except for few outliers which are very distant from the mean velocity. As extensively discussed in [10], in such situations, the total control is suboptimal because it also acts on the agents which do not need any intervention, while the sparse control strategy is locally optimal because it focuses all its strength on the small group of outliers. Such scenario is portrayed in Figure 14: Starting from the same initial datum of Figure 13, we modify the velocity of one agent so that it decisively deviates from the mean velocity. This time, the difference in the outcome of the two control strategies is much more visible. More generally, an empirical detector of configurations where it is convenient to use the

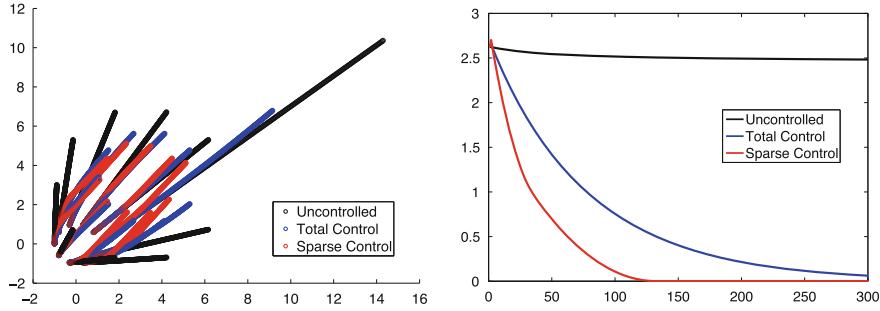


Fig. 14 Configuration with one outlier. On the left: space evolution without control (in black), with the total control (in blue), and with the sparse control strategy (in red). On the right: the respective behavior of the functional V .

sparse feedback control is the so-called *asymmetry measure*, proposed in [7, Section 3.6.5].

6 The Cucker–Dong Model

We now show how the sparse feedback control strategy previously introduced has far more reaching potential, as it can address also situations which do not match the structure (9), like the Cucker and Dong model of cohesion and avoidance introduced in [33], which is given by the following system of differential equations

$$\left\{ \begin{array}{l} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = -b_i(t)v_i(t) + \sum_{j=1}^N a(\|x_i(t) - x_j(t)\|^2)(x_j(t) - x_i(t)) + \\ \quad + \sum_{j \neq i}^N f(\|x_i(t) - x_j(t)\|^2)(x_i(t) - x_j(t)), \end{array} \right. \quad i=1, \dots, N, \quad (37)$$

The evolution is governed by an *attraction* force, modeled by a function $a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, which is, for some fixed constant $H > 0$ and $\beta \geq 0$, of the form

$$a(r) = \frac{H}{(1+r)^\beta},$$

(notice that here we have r in place of r^2 , since we write $a(\|x_i - x_j\|^2)$ in place of $a(\|x_i - x_j\|)$, and hence, a has the same form as (8)), though in general any Lipschitz continuous, nonincreasing function with maximum in $a(0)$ suffices. This

force is counteracted by a *repulsion* given by a locally Lipschitz continuous or \mathcal{C}^1 , nonincreasing function $f : (0, +\infty) \rightarrow \mathbb{R}_+$. We request that

$$\int_{\delta}^{+\infty} f(r) dr < +\infty, \quad \text{for every } \delta > 0.$$

A typical example of such a function is $f(r) = r^{-p}$ for every $p > 1$. The uniformly continuous, bounded functions $b_i : \mathbb{R}_+ \rightarrow [0, \Lambda]$, $i = 1, \dots, N$, for a given $\Lambda \geq 0$, are interpreted as a friction, which helps the system to stay confined.

It is easily seen how the above model can be rewritten as

$$\begin{cases} \dot{x}(t) = v(t), \\ \dot{v}(t) = -L(x(t))x(t) - v(t)b(t), \end{cases}$$

where for any $x \in \mathbb{R}^{dN}$, the function $L(x) \stackrel{\Delta}{=} L^a(x) - L^f(x)$ is the difference between the Laplacians of the two matrices $(a(\|x_i - x_j\|^2))_{i,j=1}^N$ and $(f(\|x_i - x_j\|^2))_{i,j=1}^N$, respectively, and we have set $vb \stackrel{\Delta}{=} (v_i b_i)_{i=1}^N$ for any $b = (b_1, \dots, b_N)$. Notice that differently from (9), now the Laplacians are acting on the variable x and not anymore on v , mixing the dynamics of the two components of the state: As a consequence, the Cucker–Dong model is a nondissipative system with singular repulsive interaction forces. Similar models considering attraction, repulsion, and other effects, such as alignment or self-drive, appear in the recent literature, and they seem effectively describing realistic situations of conditional pattern formation, see, e.g., some of the most related contributions [17, 22, 32, 41].

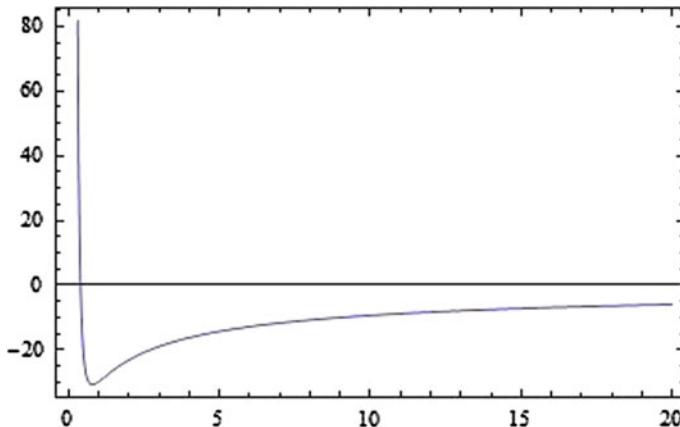


Fig. 15 Sum of the attraction and repulsion forces $h(r) = f(r) - a(r)$ as a function of the distance $r > 0$. The parameters here are $H = 50$, $\beta = 0.7$, and $p = 4$.

At first glance, it may seem perhaps a bit cumbersome to consider a rather arbitrary splitting of the force into two terms governed by the functions a and f instead of considering more naturally a unique function $h(r) \triangleq f(r) - a(r)$ of the distance $r > 0$, as depicted in Figure 15. However, as we shall clarify in short, the interplay of the polynomial decay of the function h to infinity and its singularity at 0 is fundamental in order to be able to characterize the confinement and collision avoidance of the dynamics, and such a splitting, emphasizing the individual role of these two properties, will turn out to be useful in our statements. As a matter of fact, several forces in nature do have similar behavior; for instance, the van der Waals forces are governed by Lennard-Jones potentials for which $h(r) = \sigma_f/r^{13} - \sigma_a/r^7$, for suitable positive constants σ_f and σ_a .

6.1 Pattern Formation for the Cucker–Dong Model

To quantify the behavior of the system, we introduce a quantity called the *total energy* which includes the kinetic and potential energies; for all $(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$, we define

$$E(x, v) \triangleq \sum_{i=1}^N \|v_i\|^2 + \frac{1}{2} \sum_{i < j}^N \int_0^{\|x_i - x_j\|^2} a(r) dr + \frac{1}{2} \sum_{i < j}^N \int_{\|x_i - x_j\|^2}^{+\infty} f(r) dr. \quad (38)$$

If $(x(t), v(t))$ is a point of a trajectory of system (37), we set $E(t) \triangleq E(x(t), v(t))$.

The total energy is a Lyapunov functional for system (37), and provided we are in the presence of no friction at all (i.e., $\Lambda = 0$), it is a conserved quantity.

Proposition 4 ([33, Equation (3.1)]) *For every $t \geq 0$, we have*

$$\frac{d}{dt} E(t) = -2 \sum_{i=1}^N b_i(t) \|v_i(t)\|^2.$$

Hence, if $\Lambda = 0$, then $\frac{d}{dt} E \equiv 0$.

If the attraction force at far distance is very strong (for $\beta \leq 1$), despite an initial high level of kinetic energy and of repulsion potential energy, perhaps due to a space compression of the group of particles, the dynamics is guaranteed to keep confined and collision avoiding in space at all times. If the attraction force is instead weak at far distance, i.e., $\beta > 1$, then confinement and collision avoidance turn out to be the properties of the dynamics only conditionally to *initial* low levels of kinetic energy and repulsion potential energy, meaning that the particles should not be initially too fast and too close to each other. This latter condition is formulated in terms of a total energy critical threshold

$$\vartheta \triangleq \frac{N-1}{2} \int_0^{+\infty} a(r) dr.$$

This fundamental dichotomy of the dynamics has been characterized in the following result.

Theorem 6 ([33, Theorem 2.1]) *Consider an initial datum $(x^0, v^0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ satisfying $\|x_i^0 - x_j^0\|^2 > 0$ for all $i \neq j$ and*

$$E(0) \triangleq E(x^0, v^0) < \frac{1}{2} \int_0^{+\infty} f(r) dr.$$

Then there exists a unique solution $(x(\cdot), v(\cdot))$ of system (37) with initial condition (x^0, v^0) . Moreover, if one of the two following hypotheses holds:

1. $\beta \leq 1$,
2. $\beta > 1$ and $E(0) < \vartheta$,

then the population is cohesive and collision-avoiding, i.e., there exist two constants $B_0, b_0 > 0$ such that, for all $t \geq 0$

$$b_0 \leq \|x_i(t) - x_j(t)\| \leq B_0, \quad \text{for all } 1 \leq i \neq j \leq N. \quad (39)$$

Motivated by Theorem 6, we will call *consensus region* the set

$$C \triangleq \{w \in \mathbb{R} : w \leq \vartheta\}$$

We will say that the system (37) is *in the consensus region at time t* if $E(t) \in C$. It is an obvious corollary of Theorem 6 the fact that if system (37) is in the consensus region at time T , for some $T \geq 0$, then condition (39) is fulfilled for every $t \geq T$.

Remark 9 Theorem 6 is the Cucker–Dong counterpart of Theorem 2. Indeed, for the choice of a as in (8), Theorem 2 implies that

- (i) If $\beta \leq 1/2$, then $a \notin L^1(\mathbb{R}_+)$; therefore, consensus is achieved regardless of the initial conditions and
- (ii) If $\beta > 1/2$, then $a \in L^1(\mathbb{R}_+)$, and consensus is guaranteed only if (13) is satisfied.

Remark 10 Let us stress again the fact that the word *consensus* must be intended here as a stable *cohesion and collision-avoiding* dynamics, in the spirit of the conclusion of Theorem 6. This is in contrast with the meaning of the word *consensus* in Definition 3, which describes a situation where all the agents move according to the same velocity vector. We point out that this definition of *consensus* does not imply this particular feature, but it is rather intended to make a parallel between Theorems 2 and 6, as already done by the authors in [33, Remark 1].

As for the model (7), we could construct nonconsensus events if one violates the sufficient condition (13), and also for the model (37) and in violation of the threshold $E(0) < \vartheta$, one can exhibit noncohesion events.

Example 4 (Noncohesion events [33]) Consider $N = 2$, $d = 2$, $\beta > 1$, $f \equiv 0$, $b_i \equiv 0$, and $x(t) = x_1(t) - x_2(t)$, $v(t) = v_1(t) - v_2(t)$ relative position and velocity of two agents on the line. Then we may rewrite the system as

$$\begin{cases} \dot{x} = v \\ \dot{v} = -\frac{x}{(1+x^2)^\beta}. \end{cases} \quad (40)$$

For the sake of compactness, we introduce the quantity

$$\Psi(x) \triangleq \frac{1}{(\beta-1)(1+x^2)^{\beta-1}} \quad \text{for every } x \in \mathbb{R}^2.$$

We now prove that if we are given the initial conditions $x(0) = x_0 > 0$ and $v(0) = v_0 > 0$ satisfying $v(0)^2 \geq \Psi(x(0))$, then $x(t) \rightarrow +\infty$ for $t \rightarrow +\infty$. Indeed, by direct integration in (40), one obtains $v(t)^2 = \Psi(x(t)) + v(0)^2 - \Psi(x(0))$, and it follows that $v(t) > 0$ for all $t \geq 0$. This implies that $x(\cdot)$ is increasing: Had this function an upper bound x_* , then we would have $\dot{x}(t) = v(t) \geq (\Psi(x_*) + v(0)^2 - \Psi(x(0)))^{1/2}$, which in turn implies $x(t) \rightarrow +\infty$ for $t \rightarrow +\infty$, a contradiction.

7 Sparse Control of the Cucker–Dong Model

Notice the similarity of the present situation and that of Section 5: In both cases, we have a system whose desired pattern can be enforced by decreasing a certain Lyapunov functional under the action of a sparse intervention. Given a positive constant M modeling the limited resources given to the external policy maker to influence instantaneously the dynamics, it is very natural to define the set of *admissible controls* precisely as in Definition 6: A control $u : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN}$ is admissible if it is a measurable functions, which satisfies the $\ell_1^N - \ell_2^d$ -norm constraint (33) for every $t \geq 0$. Hence, the *controlled* Cucker–Dong model is given by

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = -b_i(t)v_i(t) + \sum_{j=1}^N a \left(\|x_i(t) - x_j(t)\|^2 \right) (x_j(t) - x_i(t)) + \\ \quad + \sum_{\substack{j=1 \\ i \neq j}}^N f \left(\|x_i(t) - x_j(t)\|^2 \right) (x_i(t) - x_j(t)) + u_i(t), \end{cases} \quad i = 1, \dots, N, \quad (41)$$

where u is admissible.

The control should be exerted until $E(T) < \vartheta$ at some finite time T , and then, it should be turned off, similarly to the sparse selection of Definition 8. Since we start from $E(0) > \vartheta$, then it is necessary that our control forces the total energy to decrease, for instance by ensuring $\frac{d}{dt} E < 0$. The following technical result helps us to identify the form of admissible controls satisfying this property.

Lemma 4 *Suppose there exists a solution of the system (41). Then*

$$\frac{d}{dt} E(t) = -2 \sum_{i=1}^N b_i(t) \|v_i(t)\|^2 + 2 \sum_{i=1}^N u_i(t) \cdot v_i(t) \quad \text{for every } t \geq 0. \quad (42)$$

7.1 Extending the Sparse Control Strategy

From expression (42), it is clear that the best way our control can act on E in order to push it below the threshold is not acting on the mutual distances between agents, but according to the velocities v . Hence, we focus on the following family of controls, closely resembling (36).

Definition 9 Let $(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ and $0 \leq \varepsilon \leq M/E(0)$. We define the *sparse feedback control* $u(x, v) = (u_1(x, v), \dots, u_N(x, v))^T \in \mathbb{R}^{dN}$ associated with (x, v) as

$$u_i(x, v) \triangleq \begin{cases} -\varepsilon E(x, v) \frac{v_{\hat{i}(x, v)}}{\|v_{\hat{i}(x, v)}\|} & \text{if } i = \hat{i}(x, v), \\ 0 & \text{otherwise.} \end{cases}$$

where $\hat{i}(x, v)$ is the minimum index such that

$$\|v_{\hat{i}(x, v)}\| = \max_{1 \leq j \leq N} \|v_j\|.$$

Whenever the point (x, v) is a point of a curve $(x, v) : \mathbb{R}_+ \rightarrow \mathbb{R}^{dN} \times \mathbb{R}^{dN}$, i.e., $(x, v) = (x(t), v(t))$ for some $t \geq 0$, we will replace everywhere $u(x, v) = u(x(t), v(t))$ and $\hat{u}(x, v) = \hat{u}(x(t), v(t))$ with $u(t)$ and $\hat{u}(t)$, respectively.

Remark 11 Definition 9 makes sense if $\|v_{\hat{i}(t)}(t)\| \neq 0$ for at least almost every $t \geq 0$. Notice that if the latter condition were not holding, then $v_i(t) = 0$ for all $i = 1, \dots, N$ and for all $t \geq 0$ and hence $\dot{v}_i(t) = 0$ for all $i = 1, \dots, N$ and for all $t \geq 0$; hence, the configuration of the system would be in a steady state, and no control would be needed.

The parameter ε will help us to tune the control in order to ensure the convergence to the consensus region. Indeed, notice that if we were able to prove that $\|\bar{v}(t)\| \geq \eta$ holds for every $t \geq 0$ for some $\eta > 0$, then it would follow that

$$\frac{d}{dt} E(t) \leq 2 \left(-\varepsilon E(t) \frac{v_{\hat{i}(t)}(t)}{\|v_{\hat{i}(t)}(t)\|} \right) \cdot v_{\hat{i}(t)}(t) = -2\varepsilon E(t) \|v_{\hat{i}(t)}(t)\| \leq -2\varepsilon\eta E(t),$$

from which we obtain the estimate $E(t) \leq E(0)e^{-2\varepsilon\eta t}$ for every $t \geq 0$. Therefore, it follows that E is decreasing: This in turn implies that whenever $\varepsilon \leq M/E(0)$ holds, we have

$$\sum_{i=1}^N \|u_i(t)\| = \varepsilon E(t) \leq \frac{M}{E(0)} E(t) \leq M,$$

whence the validity of the constraint (33). Therefore, the control of Definition 9 is admissible.

By exploiting several nontrivial a priori estimates for stability (collected in [8, Section 3.2]), which were not necessary for system (9) due to its dissipative nature, we obtain the following result, which resembles closely Theorem 5.

Theorem 7 ([8, Theorem 4.1 and Proposition 4.2]) *Fix $M > 0$. Let $(x^0, v^0) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ be such that the following hold:*

- (a) $\|\bar{v}(0)\| \geq \eta > 0$;
- (b) for

$$c \triangleq \exp \left(-\frac{2\sqrt{3}}{9} \frac{M \|\bar{v}(0)\|^3}{E(0) \sqrt{E(0)} \left(\Lambda \sqrt{E(0)} + \frac{M}{N} \right)} \right)$$

it holds $c\vartheta > E(0) > \vartheta$.

Then there exist constants $T > 0$ and $\Gamma = \Gamma(x^0, v^0, \vartheta, \eta, c) > 0$, and a piecewise constant selection of the sparse feedback control of Definition 9 such that

- (1.) $\|\bar{v}(t)\| \geq \eta$ for every $t \leq T$;
- (2.) whenever $\Gamma \leq \varepsilon \leq M/E(0)$ holds, the associated absolutely continuous solution reaches the consensus region before time T .

We remark that while the stabilization of Cucker–Smale systems by means of sparse feedback controls is unconditional with respect to the initial conditions (see Theorem 5), for the Cucker–Dong model, our analysis guarantees stabilization only within certain total energy levels, which is suggesting that also stabilization can be conditional. However, the numerical experiments reported in Section 7.3 suggest that it is possible to exceed such an upper energy barrier in many cases, even if there are pathological situations for which there is no hope to steer the agents toward a cohesive configuration.

7.2 Optimality of the Sparse Feedback Control

We now pass to show that the sparse feedback control of Definition 9 is a minimizer of a variational criterion similar to (35). To this end, notice that each value of $\eta \geq 0$ appearing in Theorem 7 yields a partition of $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$ into four disjoint sets:

$$\begin{aligned}\mathcal{P}_1 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i\| < \eta\}, \\ \mathcal{P}_2 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i\| = \eta \text{ and } \exists k \geq 1 \text{ and } i_1, \dots, i_k \in \{1, \dots, N\} \text{ such that } \|v_{i_1}\| = \dots = \|v_{i_k}\| \text{ and } \|v_{i_1}\| > \|v_j\| \text{ for every } j \notin \{i_1, \dots, i_k\}\}, \\ \mathcal{P}_3 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i\| > \eta \text{ and } \exists i \in \{1, \dots, N\} \text{ such that } \|v_i\| > \|v_j\| \text{ for every } j \neq i\}, \\ \mathcal{P}_4 &\triangleq \{(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN} : \max_{1 \leq i \leq N} \|v_i\| > \eta \text{ and } \exists k > 1 \text{ and } i_1, \dots, i_k \in \{1, \dots, N\} \text{ such that } \|v_{i_1}\| = \dots = \|v_{i_k}\| \text{ and } \|v_{i_1}\| > \|v_j\| \text{ for every } j \notin \{i_1, \dots, i_k\}\},\end{aligned}$$

The above partition naturally leads to the following class of feedback controls.

Definition 10 For every $(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$, we denote with $U(x, v) \subseteq \mathbb{R}^{dN}$ the set of all vectors $u(x, v) = (u_1(x, v), \dots, u_N(x, v))^T \in \mathbb{R}^{dN}$, whose vector entries are of the form

$$u_i(x, v) = \begin{cases} -\varepsilon_i E(x, v) \frac{v_i}{\|v_i\|} & \text{if } \|v_i\| \neq 0, \\ 0 & \text{if } \|v_i\| = 0, \end{cases}$$

where the coefficients $\varepsilon_i \geq 0$ satisfy

$$\sum_{i=1}^N \varepsilon_i \leq \frac{M}{E(0)},$$

and

- if $(x, v) \in \mathcal{P}_1$ then $\varepsilon_i = 0$ for every $i = 1, \dots, N$;
- if $(x, v) \in \mathcal{P}_2$ then indicating with i_1, \dots, i_k the indexes such that $\|v_{i_1}\| = \dots = \|v_{i_k}\| = \eta$ and $\|v_{i_1}\| > \|v_j\|$ for every $j \notin \{i_1, \dots, i_k\}$, we have $\varepsilon_j = 0$ for every $j \notin \{i_1, \dots, i_k\}$;
- if $(x, v) \in \mathcal{P}_3$ then, indicating with i the only index such that $\|v_i\| > \|v_j\|$ for every $j \neq i$, we have $\varepsilon_i = M/E(0)$ and $\varepsilon_j = 0$ for every $j \neq i$;
- if $(x, v) \in \mathcal{P}_4$ then, indicating with i_1, \dots, i_k the indexes such that $\|v_{i_1}\| = \dots = \|v_{i_k}\|$ and $\|v_{i_1}\| > \|v_j\|$ for every $j \notin \{i_1, \dots, i_k\}$, we have $\varepsilon_j = 0$ for every $j \notin \{i_1, \dots, i_k\}$ and $\sum_{\ell=1}^k \varepsilon_{i_\ell} = M/E(0)$.

Remark 12 Under the hypotheses of Theorem 7, the control $u(t)$ introduced in Definition 9 belongs to $U(x(t), v(t))$ whenever $t \leq T$, since it is guaranteed that $\max_{i \leq i \leq N} \|v_i(t)\| \geq \eta$ for every $t \leq T$.

The set $U(x, v)$ is closed and convex, and, moreover, has the following very elegant alternative variational interpretation, reminiscent of Definition 7.

Proposition 5 [8, Propositions 5.2 and 5.4] For every $(x, v) \in \mathbb{R}^{dN} \times \mathbb{R}^{dN}$ and for every $M \geq 0$, set

$$m(x, v) \triangleq M \frac{E(x, v)}{E(0)} \quad \text{and} \quad K(x, v) \triangleq \left\{ u \in \mathbb{R}^{dN} : \sum_{i=1}^N \|u_i\| \leq m(x, v) \right\}.$$

Let $\mathcal{J} : \mathbb{R}^{dN} \rightarrow \mathbb{R}$ be the functional defined by

$$\mathcal{J}(u, v) \triangleq v \cdot u + \eta \sum_{i=1}^N \|u_i\|.$$

Then

$$U(x, v) = \arg \min_{u \in K(x, v)} \mathcal{J}(u, v).$$

The next result is the Cucker–Dong counterpart of Proposition 3: The sparse feedback control minimizes the decay rate of the functional E among the controls introduced in Definition 10.

Theorem 8 The feedback control of Definition 9 is an instantaneous minimizer of

$$\mathcal{D}(t, u) \triangleq \frac{d}{dt} E(t)$$

over all possible feedback controls $u \in U(x(t), v(t))$.

Similarly to what we have seen in the case of the Cucker–Smale system, the previous result shows that the most effective control strategy that the external policy maker can enact is to allocate all the resources at its disposal only on very few key agents in the system, in order to keep the dynamics bounded and collision avoiding. One of the most relevant differences with respect to Theorem 5, though, is that for the Cucker–Smale model, the stabilization can be achieved unconditionally, i.e., independently of the initial conditions (x^0, v^0) . For the Cucker–Dong model, instead, a similar sparse control strategy yields only a conditional results; i.e., we obtain stabilization conditionally to an initial energy level satisfying $\vartheta < E(0) < c\vartheta$, as stated in the condition (b) of Theorem 7. Our numerical experiments, which follow below, suggest that it is possible to exceed such an upper energy barrier, but it is unclear whether this is just a matter of fortunate choices of good initial conditions or we can actually have a broader stabilization range than the one analytically derived above.

7.3 Numerical Validation of the Sparse Control Strategy

In this section, we will report the results of significant numerical simulations on Cucker–Dong systems in the dimension $d = 2$ with and without the use of the sparse control strategy outlined in Definition 9. Throughout the section, we will keep fixed the number of agents ($N = 8$), the friction applied ($\Lambda = 0$, i.e., frictionless), and the form of the repulsive function ($f(r) = r^{-p}$). We restrict only to $N = 8$ simply for an easier visualization of the results. This means that we will vary the shape of the function a (i.e., we will act on β), the slope of the repulsion function (changing the value of p), and the maximum amount of strength of the sparse control (the parameter M). The parameter ε is always set equal to $M/E(0)$.

7.3.1 The Effect of Sparse Controls on the System

Figure 16 shows the spatial evolution and speeds of the agents of a Cucker–Dong system with $\beta = 1.1$ and $p = 2$:

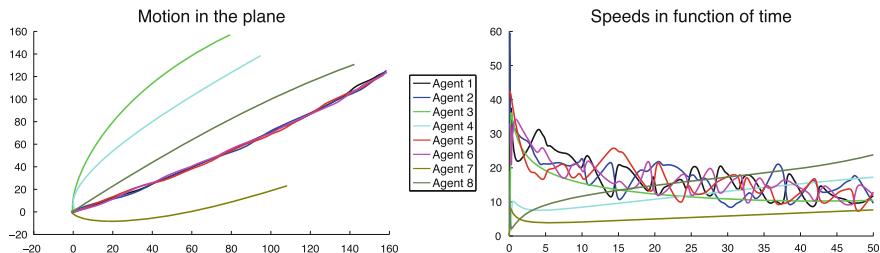


Fig. 16 Space evolution and speeds of the uncontrolled system.

Though we cannot infer the divergence of the system from this finite-time simulation, the portrayed situation seems far from going toward a flocking behavior. The only agents which seem to flock are Agent 1, Agent 2, Agent 5, and Agent 6 (black, blue, red, and magenta trajectories, respectively), as it is also visible by the corresponding speed graph, in which the speed of each agent is adjusted to the one of the other agents.

Figure 17 shows that the total energy E (the red line) is constant and far away from the consensus threshold ϑ (black line). The increase in the distances between particles is reflected in an increase in the adhesion potential energy (the one due to a , see (38)) and in a decrease in the repulsive one (due to f).

If instead we apply our sparse control strategy with $M = 35$ on the same system with the same initial conditions, the situation gets immediately far better from a consensus point of view, as Figure 18 witnesses.

The spatial evolution graph shows a braid movement which resembles a pattern near to flocking as it is commonly interpreted. The action of our control is evident from the energy profile of the system, portrayed in Figure 19, where the total energy is driven below the threshold in a very short time. The fall of the total energy is mainly due to its kinetic part (the green line), which is the only one directly affected by our control strategy. The sharp decrease of the kinetic energy is also witnessed in the graph showing the modulus of the speeds, where after a quick, strong brake at the beginning, they stabilize at a very low level.

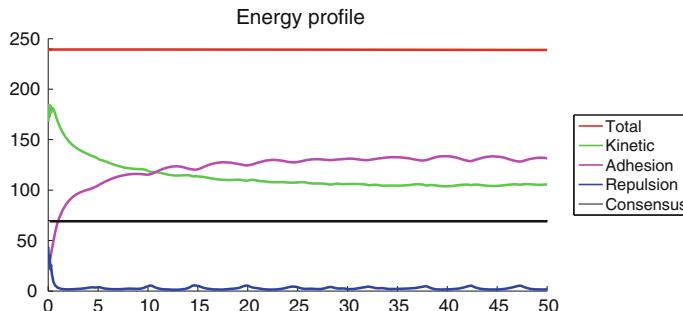


Fig. 17 Energy profile of the uncontrolled system.
Motion in the plane

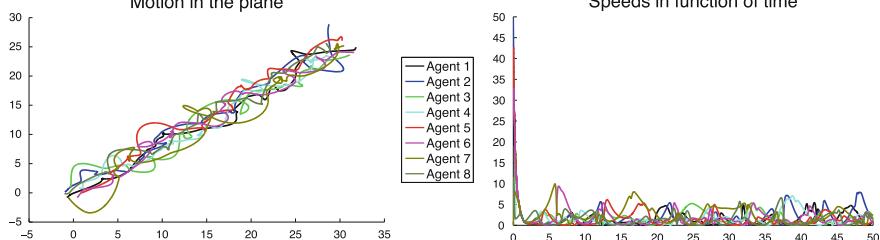


Fig. 18 Space evolution and speeds of the controlled system.

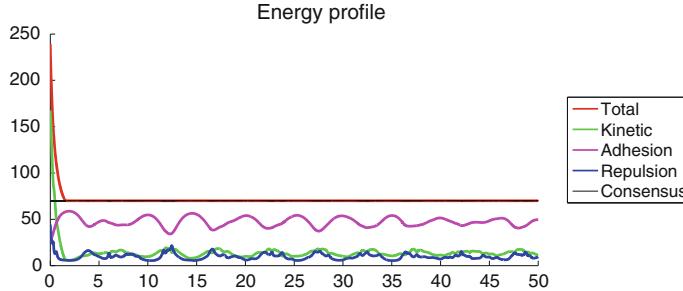


Fig. 19 Energy profile of the controlled system.

7.3.2 Tuning the Parameter M

The second case study takes into account a system with a weaker communication rate than before ($\beta = 1.02$) and with a different form of the repulsive function ($p = 1.1$), and we apply on it our control strategy with several values for M .

The top left corner of Figure 20 is the uncontrolled system: It seems legitimate to suppose that it is very unlikely that the system will converge to consensus, especially looking at its energy profile graph (top right corner of Figure 20), which shows an increase in the adhesion potential energy, phenomenon associated with an increase in the distance between particles, as already pointed out. In the second line, we see the spatial evolution graph of the same system but with the sparse control strategy acting with parameter $M = 0.1$, where the agents are starting to converge to consensus, as is also evident in their energy profile. The two bottom lines of Figure 20 display the action of controls with $M = 1$ and $M = 10$, respectively. It is clear how the situation goes better as M increases, which is due to the fact that the threshold is reached in shorter time (see the relative energy profile).

The right column of Figure 20 also clearly confirms the behavior of the decay rate of the energy as a function of M , as predicted by our analysis: $E(t)$ decreases as e^{-kM^t} , for a certain constant $k > 0$.

It is interesting to notice that convergence to the consensus region occurs even if the hypothesis (b) of Theorem 7 is not met, i.e., ϑ is very far away from $E(0)$, as it is likely to be a suboptimal sufficient condition. Indeed, in all the case studies above

$$c = \exp\left(-\frac{2\sqrt{3}}{9} \frac{M \|\bar{v}(0)\|^3}{E(0)\sqrt{E(0)} \left(\Lambda\sqrt{E(0)} + \frac{M}{N}\right)}\right) \approx 1,$$

but, nonetheless, we were able to steer the system to consensus in finite time.

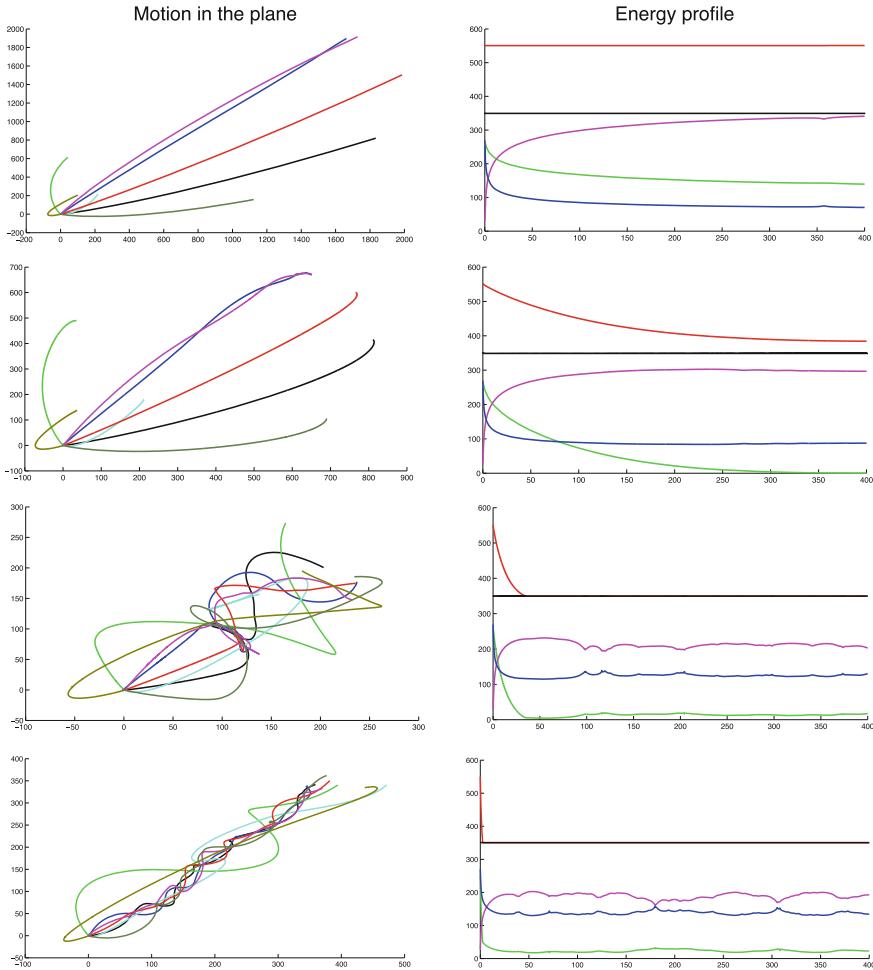


Fig. 20 Spatial evolutions (left) and relative energy profiles (right). From top to bottom: $M = 0$, $M = 0.1$, $M = 1$, $M = 10$. The colors in the left column stand for: total energy (red), consensus region (black), adhesion energy (magenta), repulsion energy (blue), kinetic energy (green).

7.3.3 A Counterexample to Unconditional Sparse Controllability

The last numerical experiment we report shows that in certain pathological situations, the sparse control strategy can fail to steer a Cucker–Dong systems to consensus.

We consider $N = 2$ agents in the dimension $d = 2$ and choose the interaction parameters as $H = 1$, $\beta = 2$, $p = 1.1$, $\Lambda = 0$, and $M = 1$. In this situation, the force balance $f - a$ is completely in favor of the repulsive force, as Figure 21 shows: This means that, regardless of the mutual positions of the agents, they shall always be repelled from each other.

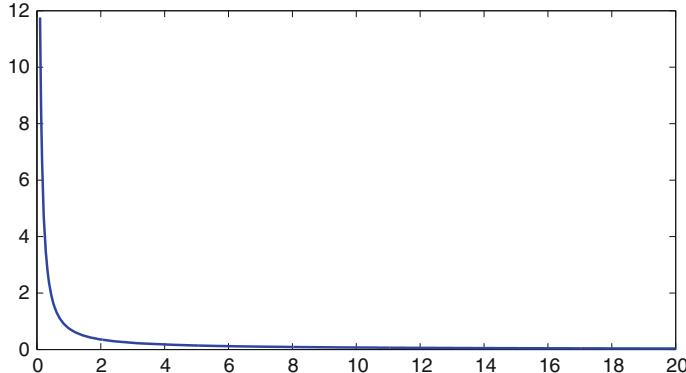


Fig. 21 Sum of the attraction and repulsion forces $h(r) = f(r) - a(r)$ as a function of the distance $r > 0$ in the case study of Section 7.3.3

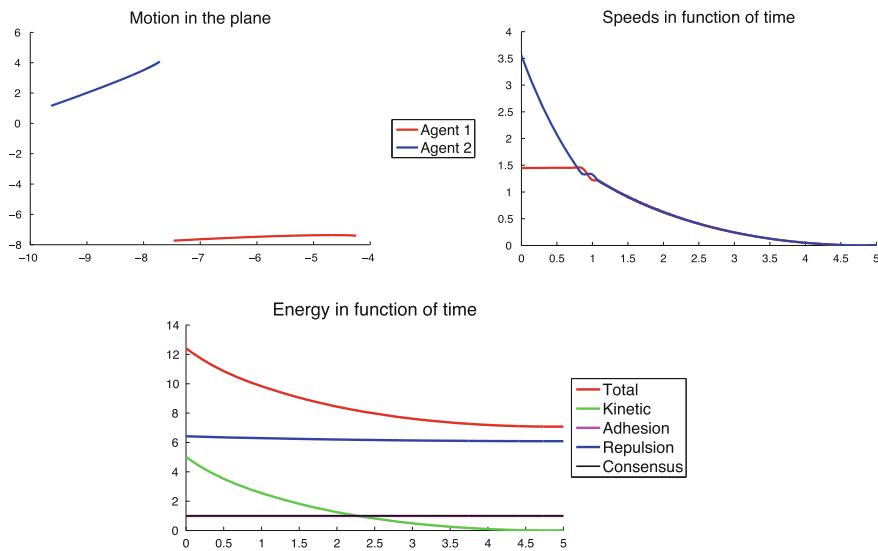


Fig. 22 Space evolution, speeds, and energy profile of the system considered in Section 7.3.3.

If we exert the sparse control strategy, the only result that we obtain is to freeze the agents where they are. Indeed, Figure 22 shows that the agents' speeds are rapidly reduced to values close to 0 as an effect of the control (also visible in the energy profile from the trajectory of the kinetic energy), but the total energy stays far away from the consensus region (the black line). The picture makes very clear that the sparse feedback control does not affect the potential energy of the system, as the sum of the adhesion and repulsion energies stays constant in time.

Furthermore, notice that as soon as we shut down the control, the two agents will start to move again in opposite directions (very slowly, since the energy of a system

without control stays constant); hence, not only the total energy remains above the threshold, but also the system is not in consensus.

However, it must be observed that the control strategy fails in this situation due to the peculiar nature of the system. As a matter of fact, being the force balance strictly repulsive, the trajectories of any solution will never remain cohesive. This leaves open the question whether there exist “nonpathological” instances of the Cucker–Dong model (in the sense that their solutions are not doomed to diverge regardless of the initial condition) for which the sparse control strategy does not work.

Acknowledgements The authors acknowledge the support of the ERC-Starting Grant “High-Dimensional Sparse Optimal Control” (HDSPCONTR - 306274).

References

1. S. M. Ahn and S.-Y. Ha. Stochastic flocking dynamics of the Cucker-Smale model with multiplicative white noises. *J. Math. Phys.*, 51(10):103301, 2010.
2. G. Albi, M. Bongini, E. Cristiani, and D. Kalise. Invisible sparse control of self-organizing agents leaving unknown environments. *To appear in SIAM J. Appl. Math.*, 2015.
3. F. Arvin, J. C. Murray, L. Shi, C. Zhang, and S. Yue. Development of an autonomous micro robot for swarm robotics. In *Proceedings of the IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 635–640. IEEE, 2014.
4. P. Bak. *How nature works: the science of self-organized criticality*. Springer Science & Business Media, 2013.
5. 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. *P. Natl. Acad. Sci. USA*, 105(4):1232–1237, 2008.
6. S. Battiston, D. Delli Gatti, M. Gallegati, B. Greenwald, and J. Stiglitz. Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk. *J. Econ. Dyn. Control*, 36(8):1121–1141, 2012.
7. M. Bongini. *Sparse Optimal Control of Multiagent Systems*. PhD thesis, Technische Universität München, 2016.
8. M. Bongini and M. Fornasier. Sparse stabilization of dynamical systems driven by attraction and avoidance forces. *Netw. Heterog. Media*, 9(1):1–31, 2014.
9. M. Bongini, M. Fornasier, F. Frölich, and L. Hagverdi. Sparse control of force field dynamics. In *International Conference on NETwork Games, COntrol and OPTimization*, October 2014.
10. M. Bongini, M. Fornasier, O. Junge, and B. Scharf. Sparse control of alignment models in high dimension. *Netw. Heterog. Media*, 10(3):647–697, 2015.
11. M. Bongini, M. Fornasier, and D. Kalise. (Un)conditional consensus emergence under perturbed and decentralized feedback controls. *Discrete Contin. Dyn. Syst.*, 35(9):4071–4094, 2015.
12. A. Borzì and S. Wongkaew. Modeling and control through leadership of a refined flocking system. *Math. Models Methods Appl. Sci.*, 25(02):255–282, 2015.
13. S. Camazine, J.-L. Deneubourg, N. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. *Self-organization in biological systems*. Princeton University Press, 2002.
14. E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.
15. 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.

16. M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and control of alignment models. *Math. Models Methods Appl. Sci.*, 25(03):521–564, 2015.
17. J. A. Carrillo, M. R. D’Orsogna, and V. Panferov. Double milling in self-propelled swarms from kinetic theory. *Kinet. Relat. Models*, 2(2):363–378, 2009.
18. J. A. Carrillo, M. Fornasier, J. Rosado, and G. Toscani. Asymptotic flocking dynamics for the kinetic Cucker-Smale model. *SIAM J. Math. Anal.*, 42(1):218–236, 2010.
19. J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil. Particle, kinetic, and hydrodynamic models of swarming. In *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, Modeling and Simulation in Science, Engineering and Technology, pages 297–336. Birkhäuser Boston, 2010.
20. J. A. Carrillo, Y.-P. Choi, and M. Hauray. The derivation of swarming models: mean-field limit and Wasserstein distances. In *Collective Dynamics from Bacteria to Crowds*, pages 1–46. Springer, 2014.
21. E. Casas, C. Clason, and K. Kunisch. Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim.*, 50(4):1735–1752, 2012.
22. Y.-L. Chuang, M. R. D’Orsogna, D. Marthaler, A. L. Bertozzi, and L. S. Chayes. State transitions and the continuum limit for a 2D interacting, self-propelled particle system. *Phys. D*, 232(1):33–47, 2007.
23. F. R. K. Chung. *Spectral graph theory*, volume 92. American Mathematical Society, 1997.
24. C. Clason and K. Kunisch. A measure space approach to optimal source placement. *Comput. Optim. Appl.*, 53(1):155–171, 2012.
25. M. A. Cohen and S. Grossberg. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Trans. Syst., Man, Cybern., Syst.*, 13(5):815–826, 1983.
26. J. Cortés and F. Bullo. Coordination and geometric optimization via distributed dynamical systems. *SIAM J. Control Optim.*, 44(5):1543–1574, 2005.
27. I. D. Couzin and N. R. Franks. Self-organized lane formation and optimized traffic flow in army ants. *P. Roy. Soc. Lond. B Bio.*, 270(1511):139–146, 2003.
28. I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433:513–516, 2005.
29. A. J. Craig and I. Flügge-Lotz. Investigation of optimal control with a minimum-fuel consumption criterion for a fourth-order plant with two control inputs; synthesis of an efficient suboptimal control. *J. Fluids Eng.*, 87(1):39–58, 1965.
30. E. Cristiani, B. Piccoli, and A. Tosin. Modeling self-organization in pedestrians and animal groups from macroscopic and microscopic viewpoints. In *Mathematical modeling of collective behavior in socio-economic and life sciences*, pages 337–364. Springer, 2010.
31. E. Cristiani, B. Piccoli, and A. Tosin. Multiscale modeling of granular flows with application to crowd dynamics. *Multiscale Model. Simul.*, 9(1):155–182, 2011.
32. F. Cucker and J.-G. Dong. A general collision-avoiding flocking framework. *IEEE Trans. Automat. Control*, 56(5):1124–1129, 2011.
33. F. Cucker and J.-G. Dong. A conditional, collision-avoiding, model for swarming. *Discrete Contin. Dynam. Systems*, 34(3):1009–1020, 2014.
34. F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Trans. Automat. Control*, 52(5):852–862, 2007.
35. F. Cucker and S. Smale. On the mathematics of emergence. *Jpn. J. Math.*, 2(1):197–227, 2007.
36. F. Cucker, S. Smale, and D. Zhou. Modeling language evolution. *Found. Comput. Math.*, 4(5):315–343, 2004.
37. S. Currarini, M. O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.
38. F. Dalmao and E. Mordecki. Cucker-Smale flocking under hierarchical leadership and random interactions. *SIAM J. Appl. Math.*, 71(4):1307–1316, 2011.
39. J. Dickinson. Animal social behavior. In *Encyclopaedia Britannica Online*. Encyclopaedia Britannica Inc., 2016.
40. D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

41. M. R. D'Orsogna, Y.-L. Chuang, A. L. Bertozzi, and L. S. Chayes. Self-propelled particles with soft-core interactions: patterns, stability, and collapse. *Phys. Rev. Lett.*, 96(10):104302, 2006.
42. Y. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans. Inform. Theory*, 56(1):505–519, 2010.
43. J. A. Fax and R. M. Murray. Information flow and cooperative control of vehicle formations. *IEEE Trans. Automat. Control*, 49(9):1465–1476, 2004.
44. A. F. Filippov. *Differential Equations with Discontinuous Righthand Sides*. Kluwer Academic Publishers, 1988.
45. M. Fornasier and H. Rauhut. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613, 2008.
46. M. Fornasier and H. Rauhut. *Handbook of Mathematical Methods in Imaging*, chapter Compressive Sensing, pages 187–228. Springer-Verlag, 2010.
47. S.-Y. Ha, J.-G. Liu, et al. A simple proof of the Cucker-Smale flocking dynamics and mean-field limit. *Commun. Math. Sci.*, 7(2):297–325, 2009.
48. S.-Y. Ha, T. Ha, and J.-H. Kim. Emergent behavior of a Cucker-Smale type particle model with nonlinear velocity couplings. *IEEE Trans. Automat. Control*, 55(7):1679–1683, 2010.
49. G. Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.
50. J. Haskovec. A note on the consensus finding problem in communication networks with switching topologies. *Appl. Anal.*, 94(5):991–998, 2015.
51. R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *J. Artif. Soc. Soc. Simulat.*, 5(3), 2002.
52. R. Herzog, G. Stadler, and G. Wachsmuth. Directional sparsity in optimal control of partial differential equations. *SIAM J. Control Optim.*, 50(2):943–963, 2012.
53. E. F. Keller and L. A. Segel. Initiation of slime mold aggregation viewed as an instability. *J. Theor. Biol.*, 26(3):399–415, 1970.
54. A. Kirman, S. Markose, S. Giansante, and P. Pin. Marginal contribution, reciprocity and equity in segregated groups: Bounded rationality and self-organization in social networks. *J. Econ. Dyn. Control*, 31(6):2085–2107, 2007.
55. A. Koch and D. White. The social lifestyle of myxobacteria. *Bioessays* 20, pages 1030–1038, 1998.
56. S. Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
57. M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.*, pages 415–444, 2001.
58. B. Mohar. The Laplacian spectrum of graphs. In Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, editors, *Graph theory, Combinatorics, and Applications*, volume 2, pages 871–898. Wiley, 1991.
59. L. Moreau. Stability of multiagent systems with time-dependent communication links. *IEEE Trans. Automat. Control*, 50(2):169–182, 2005.
60. S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Rev.*, 56(4):577–621, 2014.
61. J. F. Nash. Equilibrium points in N -person games. *Proc. Natl. Acad. Sci. USA*, 36(1):48–49, 1950.
62. H. Niwa. Self-organizing dynamic model of fish schooling. *J. Theor. Biol.*, 171:123–136, 1994.
63. F. Paganini, J. Doyle, and S. Low. Scalable laws for stable network congestion control. In *Proceedings of the 40th IEEE Conference on Decision and Control*, volume 1, pages 185–190. IEEE, 2001.
64. J. Parrish and L. Edelstein-Keshet. Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science*, 294:99–101, 1999.
65. J. Parrish, S. Viscido, and D. Gruenbaum. Self-organized fish schools: An examination of emergent properties. *Biol. Bull.*, 202:296–305, 2002.
66. L. Perea, P. Elosegui, and G. Gómez. Extension of the Cucker-Smale control law to space flight formations. *J. Guid. Control Dynam.*, 32(2):527–537, 2009.
67. B. Perthame. *Transport Equations in Biology*. Basel: Birkhäuser, 2007.

68. L. Petrovic, M. Henne, and J. Anderson. Volumetric Methods for Simulation and Rendering of Hair. Technical report, Pixar Animation Studios, 2005.
69. C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 21(4):25–34, 1987.
70. W. Romey. Individual differences make a difference in the trajectories of simulated schools of fish. *Ecol. Model.*, 92:65–77, 1996.
71. J. Shen. Cucker-Smale flocking under hierarchical leadership. *SIAM J. Appl. Math.*, 68(3):694–719, 2007.
72. M. B. Short, M. R. D’Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes. A statistical model of criminal behavior. *Math. Models Methods Appl. Sci.*, 18(suppl.):1249–1267, 2008.
73. G. Stadler. Elliptic optimal control problems with L^1 -control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44(2):159–181, 2009.
74. H. G. Tanner, A. Jadbabaie, and G. J. Pappas. Flocking in fixed and switching networks. *IEEE Trans. Automat. Control*, 52(5):863–868, 2007.
75. J. Toner and Y. Tu. Long-range order in a two-dimensional dynamical xy model: How birds fly together. *Phys. Rev. Lett.*, 75:4326–4329, 1995.
76. T. Vicsek and A. Zafeiris. Collective motion. *Phys. Rep.*, 517(3):71–140, 2012.
77. 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, 1995.
78. J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.
79. G. Wachsmuth and D. Wachsmuth. Convergence and regularization results for optimal control problems with sparsity functional. *ESAIM Control Optim. Calc. Var.*, 17(3):858–886, 2011.
80. G. Weisbuch, G. Deffuant, F. Amblard, and J.-P. Nadal. Meet, discuss, and segregate! *Complexity*, 7(3):55–63, 2002.
81. S. Wongkaew, M. Caponigro, and A. Borzì. On the control through leadership of the Hegselmann–Krause opinion formation model. *Math. Models Methods Appl. Sci.*, 25(03):565–585, 2015.
82. C. Yates, R. Erban, C. Escudero, L. Couzin, J. Buhl, L. Kevrekidis, P. Maini, and D. Sumpter. Inherent noise can facilitate coherence in collective swarm motion. *Proceedings of the National Academy of Sciences*, 106:5464–5469, 2009.

A Kinetic Theory Approach to the Modeling of Complex Living Systems

Diletta Burini, Livio Gibelli and Nisrine Outada

Abstract In this chapter, a mathematical structure is derived to provide a general framework toward the modeling of space-homogeneous living systems, according to the kinetic theory for active particles. This structure can be adapted to study a variety of processes such as collective learning and social dynamics. Simple models regarding learning in a classroom and the dynamics of the criminality are presented to illustrate how the general modeling strategy operates in well-defined applications. Future research directions using the proposed approach are discussed.

1 Introduction

Living systems are relevant example of complex systems, namely systems composed of many interacting entities whose collective dynamics is more, and different, than the sum of individual behaviors [5]. Indeed, such systems show collective emerging behaviors generated by a kind of swarming intelligence which involves all the interacting entities [11, 24, 30]. It is worth noticing that at a qualitative level, emerging behaviors are often reproduced under suitable input conditions, though quantitative matches with the observations are rarely obtained. In fact, small changes in the input conditions often generate large deviations. In some cases, these break out the macroscopic (qualitative) features of the collective emerging dynamics and lead to highly unpredictable events with dramatic consequences. One of such events, which is of

D. Burini · L. Gibelli (✉)

Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy
e-mail: livio.gibelli@polito.it

D. Burini
e-mail: diletta.burini@polito.it

N. Outada
Faculté des Sciences Semlalia, Marrakesh, Morocco
e-mail: outada@ljll.math.upmc.fr

N. Outada
Laboratoire Jacques Louis-Lions, Université Pierre et Marie Curie, Paris, France

paramount interest in social sciences, in general, and in economics in particular, is the so-called Black Swan, namely an extreme event, largely unpredictable at a collective level, originating from apparently rational and controlled individual behaviors [75].

Most of the mathematical models of living systems are based on the game-theoretic framework. In its original formulation, game theory was used to analyze situations of conflict or cooperation in economics [62, 65]. Two major streams of extensions of classical game theory have been developed as early as in the sixties of the last century, namely differential games [71] and games with incomplete information [47]. The former are characterized by continuously time-varying strategies and payoffs with a dynamical system governed by ordinary differential equations, while the latter players are supposed to ignore information about the other players for what strategies and payoffs are concerned.

An important breakthrough of game theory is behavioral game theory, also called experimental game theory [28, 45]. The main idea is that instead of analyzing games theoretically, experimenters get real people to play them and record results. The remarkable finding is that in many situations, people react instinctively and play according to heuristic rules and social norms rather than adopting the strategies indicated by rational game theory.

Evolutionary game theory is an extension of the classical paradigm toward this concept of bounded rationality. Unlike classical game theory, the focus is on large populations of individuals who interact at random rather than on a small number of individuals who play a repeated game and individuals are assumed to employ adaptive rules rather than being perfectly rational. These rules change in time depending on how much players take into consideration the earlier history of the game and how long would think ahead. A static, equilibrium-based analysis turns out to be insufficient to provide enough insight into the long-run behavior of such systems [51]. It is worth noticing that evolutionary game theory overlaps with another active area of research in mathematical biology, namely population ecology and population genetics [79] as explicitly discussed in Reference [51].

Although behavioral rules control the system on the microscopic, individual level, evolutionary game theory in its standard form considers population dynamics on the aggregate level. However, in practical applications, it may be important to properly account for individual preferences, payoffs, strategy options (heterogeneity in agent types), and/or specific local connection among individuals (structural heterogeneity) [74]. The systematic investigation of such heterogeneities requires a change of perspective in the description of the system from the aggregate level to the individual level. The resulting huge increase in the relevant system variables makes most standard analytical techniques, based on differential equations, fixed points, etc., largely inapplicable and prompted the development of agent-based models. The latter are computational models which simulate the actions and interactions of autonomous agents (both individual or collective entities such as organizations or groups) with the aim to assess the collective emerging behavior.

A trend is emerging which tries to provide a unified approach, suitable to link methods of statistical mechanics and game theory. A first approach is due to Helbing, who has the merit of pointing out that individual interactions need to be modeled

by methods of game theory [48, 49]. Another approach is provided by the mean field games, where interactions among individuals are evaluated by supposing that each individual contributes to the creation of a mean field which then acts back on each individual [56, 57]. Both approaches mainly focus on the continuum limit of infinitely many individuals. The strong similarity between the mathematics of statistical physics and games where players have limited rationality has been recently pointed out [81].

The approach based on the *kinetic theory of active particles*, which is the central topic of this chapter, encompasses and extends the mathematical tools of the various approaches aforementioned. As such, it represents, to the authors' opinion, the most promising approach to the modeling of living systems.

As in game theoretical models, the evolution of the system depends upon a game strategy, which determines the individual response to interactive encounters. At the same time, the system heterogeneity is fully accounted for in a more refined way than in the agent-based framework since mathematical models are generally stated in terms of integro-differential equations amenable to mathematical and computational analysis.

The mathematical framework has been initially formalized in the book [10] and subsequently developed in a sequel of papers as reviewed in [13]. Although interactions are always modeled by theoretical tools of evolutionary game theory, in [10], their output is conditioned on the state of interacting pairs, while in [13], it also depends on the overall probability state of the system. Arguably, further developments are expected in the next years spurred by applications in various fields of applied and life sciences.

The contents of the rest of the chapter are structured as follows.

Section 2 provides a phenomenological description of living systems and illustrates the strategy which is here adopted for a consistent modeling of them.

Section 3 presents a general mathematical structure that retains complexity features of living systems. This structure is in the form of a system of integro-differential equations, and it is deemed to provide the framework for the derivation of specific models.

Section 4 outlines the numerical methods that can be used to numerically solve this class of equations.

Sections 5 and 6 introduce two topics which are currently under intense research, namely collective learning dynamics and social dynamics. Specific applications are then selected to enlighten how the mathematical structure proposed in this chapter can be applied to the modeling of systems constituted of many interacting entities.

Section 7 presents a personal quest toward some research perspectives selected according to our own bias to address future research in the challenging field of mathematics combined with the soft sciences.

2 Complexity Features of Living Systems

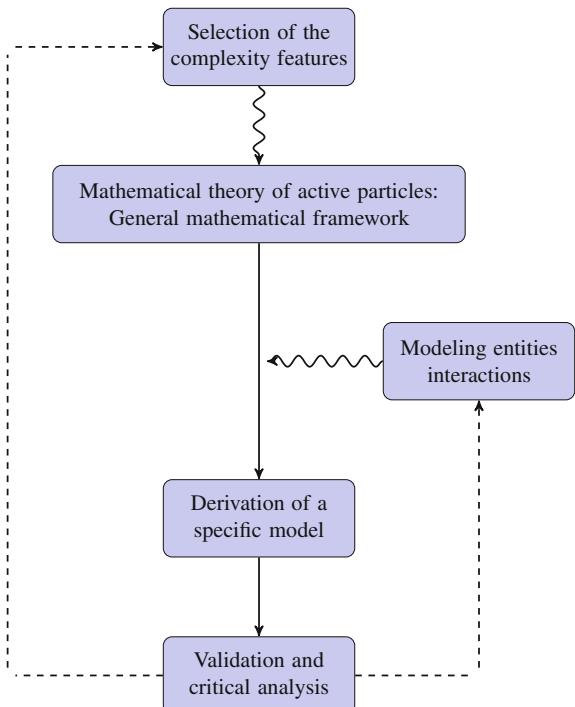
The main challenge in modeling living systems is that unlike inert matter, the *active* ability of individuals to develop behavioral strategies and to adapt them to the context makes observable effects arise from often non-evident causes. A direct consequence is that similar causes can have different effects and different causes similar effects. Furthermore, small changes of causes may have large effects, whereas large changes can also only result in small effects.

The lack of background field theories has been well highlighted in biology [60] and in the theory of evolution [61] and explains why, despite a long-lasting interaction and successful results, mathematics is still very far from understanding the complexity of biological phenomena.

As discussed in [13], in order to overcome this difficulty, a proper strategy needs to be adopted in the modeling of living systems. This strategy is depicted in Figure 1 and can be summarized as follows:

- Selection of the main complexity features of living systems under consideration.
- Derivation of a general mathematical structure suitable to capture the said complexity features.

Fig. 1 Sketch of the modeling strategy. A general mathematical structure is derived that captures some selected complexity features of living systems. Specific models can be derived by properly modeling interactions at the microscale. Validation against empirical data permits to critically assess both the selection of the complexity features and the modeling of entities' interactions.



- Derivation of mathematical models of specific systems from the general mathematical structure by modeling interactions at the scale of individuals using stochastic evolutive game theory and individual/collective learning theory.
- Validation of models based not only on their ability to reproduce available empirical data at a quantitative level, but also qualitative observed emerging behaviors.
- Critical analysis of the selection of the complexity features and the modeling of entities' interactions.

According to the critical analysis presented in [13], the main complexity features of living systems are the following:

- *Ability to express a strategy*: Living entities are able to develop specific strategies to fulfill their goals, typically their own well-being, depending on their own state and on that of the entities in the surrounding environment.
- *Heterogeneity*: The ability to express a strategy is heterogeneously distributed. Heterogeneous behaviors can play an important role in determining the overall collective dynamics, and whenever irrational behaviors even of a few entities appear, large deviations from the rationality-driven dynamics can be observed.
- *Nonlinear interactions*: Interactions are nonlinearly additive and involve immediate neighbors, but in some cases also distant entities due to the ability of living systems to communicate. As a consequence, the global action exerted on an entity by a group of others does not consist merely in the linear superposition of the actions exerted individually by them.
- *Learning and adaptation*: The strategic ability of entities and the features of their mutual interactions evolve in time, due to the ability to learn from past experience and to adapt to the time-changing external environment.
- *Selection and evolution*: New groups and/or new individual entities may form that are more suited to the evolving environment. In the course of time, some of such groups or individual entities may disappear and new ones become dominant.

As it will be discussed in detail in the following section, the kinetic theory of active particles has the capability of accounting for all these features. According to such a theory, living entities are regarded as particles and their ability to express a strategy is taken into account by defining their microscopic state through a specific additional variable. The heterogeneous behavior is dealt with by describing the system by means of a probability distribution over the microscopic state. Interactions are modeled by stochastic evolutionary games in such a way that the learning and adaptation features of living systems can be fully accounted for as well as the onset and disappearance of groups of individuals are more suited to the evolving environment.

3 A Mathematical Structure Toward Living Systems

The entities that comprise the system are referred to as *active particles* and are assumed to be distributed in a network of nodes labeled by the subscript $i = 1, \dots, n$.

Active particles are grouped into *functional subsystems* (FSs), labeled by the subscript $j = 1, \dots, m$, so that active particles within a FS share the same strategy. Entities' heterogeneity is accounted for by means of a microscopic state variable which is referred to as *activity*, u . Hence, the state of each FS is defined by a probability distribution over the said activity variable

$$f_{ij}(t, u) : [0, T] \times D_u \rightarrow \mathbb{R}_+, \quad (1)$$

where the domain of definition of the activity, D_u , is assumed to be $[-1, 1]$ or the whole real axis for probability distributions which decay to zero at infinity.

Moments provide a description of the system at the macroscopic scale. In detail, the zeroth order moment gives the number density

$$n_{ij}[f_{ij}](t) = \mathbb{E}_{ij}^0[f_{ij}](t) = \int_{D_u} f_{ij}(t, u) du, \quad (2)$$

which provides the number of active particles in the i th node and the j th FS. Higher order moments correspond to additional macroscopic variables and are defined as

$$\mathbb{E}_{ij}^p[f_{ij}](t) = \frac{1}{n_{ij}[f_{ij}](t)} \int_{D_u} u^p f_{ij}(t, u) du, \quad p = 1, 2, \dots \quad (3)$$

Active particles interact within the same functional subsystem as well as with particles of other subsystems and are acted upon by external actions. Interactions are thus grouped into two categories:

- *Microscopic scale interactions*: These include individual-based interactions between particles belonging to the same or to different FSs and external actions.
- *Microscopic–macroscopic scale interactions*: Interactions between particles and FSs viewed as a whole being represented by their mean value.

As summarized in Figure 2, in the microscopic scale interactions, active particles can be distinguished according to their different roles. More specifically, a *candidate particle* of the p th node and q th FS with activity u_* can gain, in probability, the state u of the *test particle* of the i th node and j th FS as a consequence of the interaction with the *field particle* of h th node and k th FS with activity u^* . Hereinafter, the term ij -particle is used to denote a particle in the i th node and j th FS.

The mathematical framework describing different kinds of interactions will be specified, for each type of interactions, by means of two terms. The *interaction rate* provides the frequency of interactions, while the *transition probability density* describes the probability density that a candidate pq -particle falls into the state of the test ij -particle after the interaction with a field hk -particle.

The evolution of the probability distribution is obtained by a balance of particles within elementary volumes of the space of microscopic states, the inflow and outflow of particles being related to the aforementioned interactions. Detailed calculations are not reported here as the interested reader can find them in some recent papers

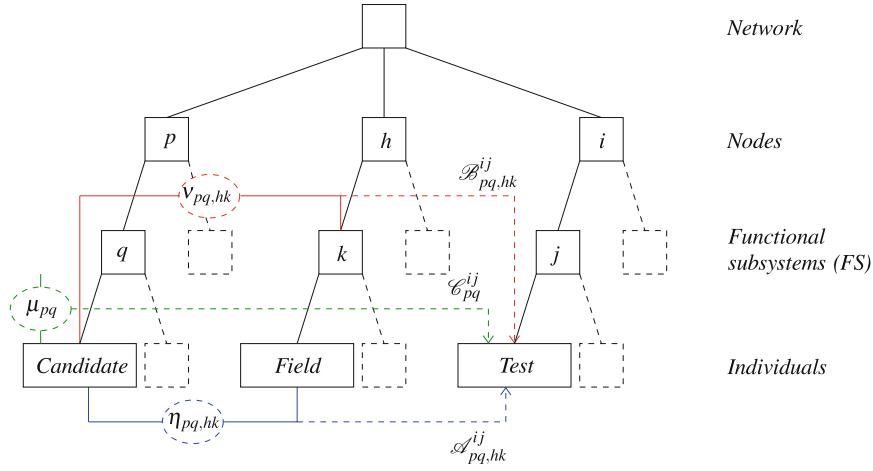


Fig. 2 A complex living system can be represented as a network of interconnected nodes. These, in turn, comprise several FSs in which individuals are grouped in accordance with their strategy. Interactions can be classified in microscopic and microscopic–macroscopic interactions. The former occur between individuals (blue color) and individuals and external fields (green color), while the latter occur between individuals and FSs viewed as a whole being represented by their mean value (red color). In microscopic interactions, candidate individuals interacting with field individuals become test individuals.

such as [13, 33]. The result is the following structure:

$$\partial_t f_{ij}(t, u) = A_{ij}[\mathbf{f}](t, u) + B_{ij}[\mathbf{f}](t, u) + C_{ij}[\mathbf{f}, \varphi](t, u), \quad (4)$$

where \mathbf{f} denotes the set of all distribution functions and

$$\begin{aligned} A_{ij}[\mathbf{f}](t, u) &= \sum_{p,h=1}^n \sum_{q,k=1}^m \int_{D_u \times D_u} \eta_{pq,hk}[\mathbf{f}](u_*, u^*) \mathcal{A}_{pq,hk}^{ij}[\mathbf{f}](u_* \rightarrow u | u_*, u^*) \\ &\quad \times f_{pq}(t, u_*) f_{hk}(t, u^*) du_* du^* \\ &\quad - f_{ij}(t, u) \sum_{h=1}^m \sum_{k=1}^n \int_{D_u} \eta_{ij,hk}[\mathbf{f}](u, u^*) f_{hk}(t, u^*) du^*, \end{aligned} \quad (5)$$

$$\begin{aligned} B_{ij}[\mathbf{f}](t, u) &= \sum_{p,h=1}^n \sum_{q,k=1}^m \int_{D_u} v_{pq,hk}[\mathbf{f}](u_*) \mathcal{B}_{pq,hk}^{ij}[\mathbf{f}](u_* \rightarrow u | u_*, \mathbb{E}_{hk}^1[f_{hk}]) f_{pq}(t, u_*) du_* \\ &\quad - f_{ij}(t, u) \sum_{h=1}^m \sum_{k=1}^n v_{ij,hk}[\mathbf{f}](u), \end{aligned} \quad (6)$$

and

$$\begin{aligned} C_{ij}[\mathbf{f}, \varphi](t, u) &= \sum_{p=1}^n \sum_{q=1}^m \int_{D_u \times D_u} \mu_{pq}[\mathbf{f}, \varphi](u_*, u^*) \mathcal{C}_{pq}^{ij}[\mathbf{f}, \varphi](u_* \rightarrow u | u_*, u^*) \\ &\quad \times f_{pq}(t, u_*) \varphi_{pq}(t, u^*) du_* du^* \\ &\quad - f_{ij}(t, u) \int_{D_u} \mu_{ij}[\mathbf{f}, \varphi](u, u^*) \varphi_{ij}(t, u^*) du^*. \end{aligned} \quad (7)$$

In Eqs. (5)–(7), square brackets have been used to denote the dependence on the distribution functions which highlight the nonlinear nature of interactions. Interaction rates and the transition probability densities have been denoted by $\eta_{pq,hk}$, $v_{pq,hk}$, μ_{pq} and $\mathcal{A}_{pq,hk}^{ij}$, $\mathcal{B}_{pq,hk}^{ij}$, \mathcal{C}_{pq}^{ij} , respectively, while external actions are assumed to be known functions of time and of the activity variable:

$$\varepsilon_{ij}(t, u) = \varepsilon_{ij}(t) \psi_{ij}(u) : [0, T] \times D_u \rightarrow \mathbb{R}_+, \quad (8)$$

where ε_{ij} models the intensity of the action in the time interval $[0, T]$, while ψ_{ij} models the intensity of the action over the activity variable.

It is worth noticing that alternative mathematical structures can be used. As an example, external actions could be accounted for as follows:

$$\partial_t f_{ij}(t, u) + \partial_u [\varphi_{ij}(t, u) f_{ij}(t, u)] = A_{ij}[\mathbf{f}](t, u) + B_{ij}[\mathbf{f}](t, u). \quad (9)$$

Remark 1 According to the conjecture proposed in [7] and formalized in [11], the interaction domain D_u is assumed, for simplicity, constant. However, a dependence of D_u on the distribution function should be taken into account that interactions occur in a sensitivity domain $\mathcal{Q} \subseteq D_u$ defined so as to contain a critical number n_c of field particles. Integration of the distribution function over the activity variable yields the topological domain of interaction $\mathcal{Q} = [u - s_m[f; n_c], u + s_M[f; n_c]]$, where $s_m, s_M > 0$. However, the solution is unique only in some special cases. For instance, when u is a scalar defined over the whole real axis, the sensitivity is symmetric with respect to u .

Remark 2 The mathematical structure presented in this section includes only interactions which preserve the number of particles. This structure provides the framework for the derivation of the class of models presented in the next two sections. As it is shown in [15], a technical generalization can include also proliferative/destructive interactions and a Darwinian type dynamics.

Let us now consider the problem of *modeling interactions* at the microscopic scale to be inserted into Eqs. (5)–(7) to obtain specific models. Various recent papers have contributed to this topic. For instance, [12, 35] have shown how interactions can be modeled by games, where the output of the interactions is conditioned not only by the state of the interacting entities, but also by the probability distribution over such states. Models which account for micro–macro–interactions have been also proposed

for the modeling of migration phenomena in [52] and opinion formation in small networks in [53].

Bearing all above in mind, let us consider, separately, the modeling of the interaction rates and the transition probability densities.

Interaction rate: The modeling of interactions requires the definition of different concepts of *distance* between interacting entities. In detail:

- *Microstate–microstate distance* involves the microscopic states u_* and u^* of the candidate/test and field particles, respectively, and can be defined as $|u_* - u^*|$.
- *Microstate–macro-state distance* involves the microscopic state u_* of the candidate particle and mean value of the activity \mathbb{E}^1 of the group of particles belonging to a functional subsystem and can be defined as $|u_* - \mathbb{E}^1|$.
- *Affinity distance* refers to the interaction between active particles characterized by different distribution functions. This distance is introduced according to the general idea that two systems with close distributions are *affine* and can be defined as by $\|f_{pq} - f_{hk}\|$, where $\|\cdot\|$ is a suitable norm to be chosen depending on the physics of the system under consideration.
- *Hierarchic distance* refers to the interaction between active particles which belong to different FSs. This distance is defined by a function $h(j, l)$, to be specified, that weights the relative influence of the l th FS on the j th FS.

The overall distance, which one can refer to as *social metrics*, is a weighted sum of all these distances. In general, interactions decay with the distance, where heuristic assumptions lead to a decay described by exponential terms or rational fractions.

Transition probability density: The dynamics of interactions can be modeled by theoretical tools from evolutionary and behavioral game theory [28, 46, 66], which provides features to be introduced into the general mathematical structure in order to obtain specific models. Additional tools are provided by learning theory within the broader context of statistical dynamics and probability theory [26]. For instance, the following types of games are often used:

- *Competition (dissent):* The interacting particle with higher status increases its status by taking advantage of the other with lower status. Therefore, the competition is advantageous for only one of the two players involved in the game.
- *Cooperation (consensus):* The interacting particles tend to share their microscopic states by decreasing the difference between their states due to a sort of attraction effect [50, 63].
- *Learning:* One of the two particles modifies, independently of the other, its microscopic state, while the other reduces the distance by a learning process.
- *Hiding/chasing:* One of the two particles attempts to increase its distance from the state of the other one (hiding), which conversely tries to reduce it (chasing).

In general, all aforesaid types of games can occur simultaneously. In some cases, their occurrence is ruled by a threshold on the distance between the states of the

interacting particles [12]. Such a threshold may be assumed to be a constant value even though, as pointed out in a recent paper [35], its dependence on the state of the system should not be neglected since it has an important influence on the overall dynamics.

4 Computational Methods

The development of the numerical methods for solving the class of equations proposed in this chapter needs to take into account that the dependent variables are probability distributions depending on time. Deterministic methods of solution approximate the said distribution at a number of collocation points over the activity variable [34]. This transforms the original integro-differential system into a coupled system of ordinary differential equations which correspond to the discrete values of the probability distributions in each collocation point. Classical numerical techniques for solving ordinary differential equations can then be employed. It should be noted that the resulting system of ordinary differential equations can be usually very large for a network where in each node, various functional subsystems interact. Although parallel computing can be used to alleviate the computational burden [40, 41], alternative numerical approaches are certainly of interest.

Monte Carlo particle simulation methods turn out to be ideally suited for solving the class of equations under consideration [68]. They originate from the Direct Simulation Monte Carlo (DSMC) scheme proposed by G.A. Bird [22] and are by far the most popular and widely used simulation methods in rarefied gas dynamics with applications ranging from vacuum technology [42] to dense fluids [9].

The basic idea consists in representing the distribution function by a number of computational particles. These particles move through the network, unless migration dynamics is disregarded, and interact according to stochastic rules derived from the governing kinetic equation. Moments are obtained through weighted averages of the particle properties. Compared to deterministic methods of solution, the application of stochastic methods shows some important advantages such as computational efficiency and ability to easily account for sophisticated individual decision processes.

A distinction is sometimes made between a Monte Carlo simulation and method, the former being the direct coding of a natural stochastic process, while the latter the solution by probabilistic methods of non-probabilistic problems. However, in many cases, such a distinction cannot be maintained and the same computer code can be regarded simultaneously as a direct physical transcription of the system dynamics and as a stochastic solution of kinetic equation. This is the case of DSMC which was introduced based on physical reasoning, but it has been later proved to converge, in a suitable limit, to the solution of the Boltzmann equation [78].

Numerical solutions of the case studies reported in Subsections 5.1 and 6.1 have been obtained by means of a stochastic and a deterministic methods, respectively.

5 Modeling Learning Dynamics

A rough simplification of the learning dynamics in complex living systems identifies three sequential steps. The first step is the *perception*. Each individual possesses a perception domain, that is, a domain within which the presence of other individuals is felt with a different intensity that depends on the mutual “distance” to be properly defined. Then, *interactions* take place and trigger a *learning process* which modifies their level of knowledge.

In the pioneering studies of the 1950s of the last century, two main complementary schools can be identified. The first developed *stochastic learning models* where the subject of the experiment receives a stimulus, and then, he makes one of a number of possible responses and, accordingly, receives either a reward or a punishment [27]. It is assumed that the responses have a probability of occurrence which evolve in time depending on the outcome of each trial. The learning process consists in changing the probability of the responses and the rules that modify them. The second school developed *stimulus sampling models* where the subject in a learning experiment samples a population of stimuli, or “cues,” on each trial and his probability of making a given response depends on the proportion of sampled stimuli that are “conditioned” or “connected” to the response [37].

In more recent years, several studies have been devoted to the search of mathematical algorithms which could approximate the “knowledge” of some unknown information. In this respect, in Reference [31], the authors present an interesting survey which characterizes the mathematical theory of learning as a method to approximate systems described by a large number of variables. In this respect, two fundamental types of learning can be distinguished, namely *cognitive* and *social* learning. The former refers to the case when an individual increases its mental knowledge, while the latter occurs when the individual learns new behaviors from others, e.g., language acquisition by children and behavioral learning in the animal kingdom. In such context, one of the great contributions to biology is the replacement of the concept of *typological thinking* by that *population thinking*, proposed by Mayr [61], linked to the concept of mutations and selection can explain various aspects of the theory of evolution.

Mayr’s theory has strongly motivated in later years the development of the evolutionary game theory set up by Sigmund [51]. This latter has found many applications in non-biological fields like economics or learning theory and presents an important enrichment of “classical” game theory, which is centered on the concept of a rational individual. In contrast, evolutionary game theory deals with entire populations of players, all programmed to use some strategy (or type of behavior). Strategies with high payoff will spread within the population (e.g., this can be achieved by learning). The payoffs depend on the actions of the coplayers and hence on the frequencies of the strategies within the population. Since these frequencies change according to the payoffs, this yields a feedback loop. This dynamics is the object of evolutionary game theory and the growing interest toward living, hence complex, systems composed

of many interacting agents led to an important field of application of the learning phenomena, offered by the stochastic evolutionary game theory [4, 66].

Furthermore, a vast literature was developed on the cognitive and social learning for different types of living systems, with different modeling strategy due to the specific learning process in consideration, e.g., crowds, swarms, and schools [11], intelligent robots, and social dynamics. In opinion dynamics, various approaches have been used to investigate the different mechanisms leading to group polarization and to choice averaging [44]. Some of them are inspired by the kinetic theory of rarefied gas, supported by efficient simulation algorithms based on Monte Carlo methods [68]. More complex learning dynamics include learning with Darwinian mutations and adaptation where the process can generate entities with higher ability as a consequence of an evolution process [61] and also multiple learning of different abilities, e.g., opinion formation linked to welfare dynamics, as discussed in [12], where it is shown that different dynamics can lead to unpredictable events [75]. Moreover, many studies in the modern world are focused more on issues than on disciplines. As such, they are at the intersection of different fields of knowledge such as genetics and computer science, cognitive science and neuroscience, economics and behavioral science or art, and computers.

Lastly, the paper [26] presents an overview and critical analysis of the literature on the modeling of learning dynamics. The authors propose their own approach, based on suitable development of methods of the kinetic theory and theoretical tools of evolutionary game theory, with the objective of developing a mathematical theory of perception and learning in view of their application to modeling complex systems, which can develop a collective intelligence [8]. Their formulation can be viewed as an extension of the concept of population thinking and of the theory of evolution, and the important motivations to the contents of this chapter are induced by the idea that the mathematical structure might include features which could make it interesting in different fields of life sciences.

In the next section, it is shown how the learning process that takes place in a classroom can be modeled according to the kinetic theory of active particles.

5.1 Case Study: Learning in a Classroom

Early studies on these processes have been conducted by psychologists and sociologists [69, 77]. Recent thought on educational psychology suggests that learning processes take place while people participate within social communities [59]. Also, learning is not restricted to any type of intentional education but involves all kinds of social activities [18]. Within a more restricted social context, teaching–learning processes that take place in the classroom are being extensively investigated. Therefore, this novel and multidisciplinary approach links psychological and sociological theories of impact [58], education psychology [18], mathematics, computer science, and statistical physics [21].

The case study presented in the following has been developed in [26]. It discusses the learning process that takes place in a classroom and, more specifically, how students' performances are affected by group studying.

The rest of the subsection is organized into two parts. The first part develops the mathematical formulation of the test case, while the second part briefly presents and discusses the numerical results.

5.1.1 Mathematical Formulation

The students are assumed to attend lectures given by a qualified teacher and to interact with each other within working groups. Following [25], the influence of the group structure on the learning process is assessed by subdividing the students into two different sets, i.e., subsystems, namely “high achieving” (HA) and “low achieving” (LA). Space dynamics is not accounted for, and therefore, the system is supposed to comprise only one node. The teacher is not included in the description of the system, but his presence is explicitly included in the model through an interaction term which describes his contribution to the student learning process. The activity variable for each of them expresses their own level of knowledge.

The application uses a simple particularization of the structure (4)–(7). More in detail:

$$\partial_t f_j(t, u) = A_j[\mathbf{f}](t, u) + C_j[\mathbf{f}](t, u), \quad (10)$$

where

$$\begin{aligned} A_j[\mathbf{f}](t, u) &= \sum_{k=1}^2 \int \int_{D_u \times D_u} \eta_{j,k}(u_*, u^*) \mathcal{A}_{j,k}^j(u_* \rightarrow u | u_*, u^*) f_j(t, u_*) f_k(t, u^*) du_* du^* \\ &\quad - f_j(t, u) \sum_{k=1}^2 \int_{D_u} \eta_{j,k}(u, u^*) f_k(t, u^*) du^*, \end{aligned} \quad (11)$$

$$\begin{aligned} C_j[\mathbf{f}](t, u) &= \int \int_{D_u \times D_u} \mu_j(u_*) \mathcal{C}_j^j(u_* \rightarrow u | u_*, u^* = \bar{u}) f_j(t, u_*) \varphi(t, u^*) du_* du^* \\ &\quad - f_j(t, u) \mu_j(u) \int_{D_u} \varphi(t, u^*) du^*. \end{aligned} \quad (12)$$

We take the external action of the teacher, whose level of knowledge has been denoted by \bar{u} , as follows

$$\varphi(t, u^*) = \delta(u^* - \bar{u}), \quad (13)$$

where $\delta(\cdot)$ is the Dirac delta function.

Three different cases are considered, which are denoted by (a), (b), and (c):

- (a): $\partial_t f_j(t, u) = C_j[\mathbf{f}](t, u)$,
- (b): $\partial_t f_j(t, u) = A_j[\mathbf{f}](t, u) + C_j[\mathbf{f}](t, u)$ with $k = j$,
- (c): $\partial_t f_j(t, u) = A_j[\mathbf{f}](t, u) + C_j[\mathbf{f}](t, u)$.

Case (a) corresponds to the traditional teaching approach; that is, students are supposed to only attend lectures of the teacher. In cases (b) and (c), the students are also engaged in collaborative work forming groups of two individuals. More specifically, in case (b), the groups are homogeneous and the members of each group are selected among students having similar initial achievements, while in case (c), the groups are heterogeneous and the members are chosen at random. It is worth noticing that the interaction term which accounts for student–teacher interactions is linear in the distribution function.

Interaction rates and transition probability densities for all three cases are summarized in Fig. 3 and are discussed in some more depth in the following. For more details, the reader is referred to [26].

Interaction rates: Only microscopic scale interactions are considered, and the time scaling has been chosen so that the interaction rates $\eta_{j,k}$ and μ_j are equal to one.

Transition Probability Densities:

For the student–teacher interaction

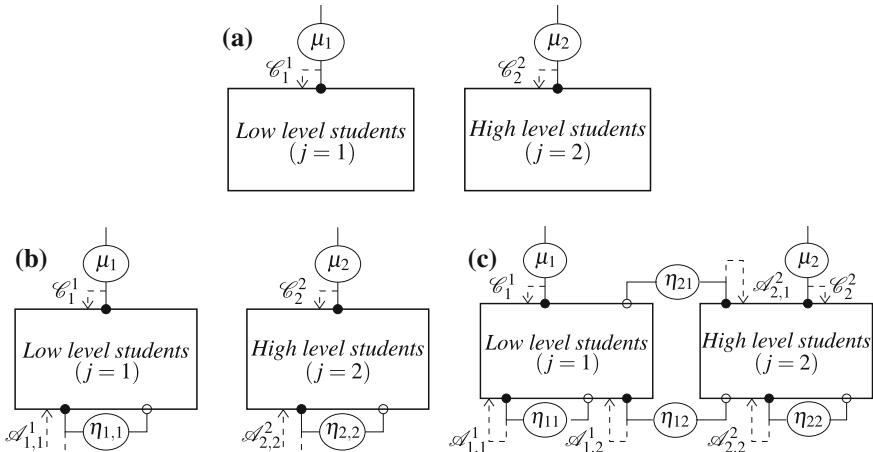


Fig. 3 Interaction rates, $\eta_{q,k}$ and μ_q and transition probability densities $\mathcal{A}_{p,q}^j$ and \mathcal{C}_p^j . Solid circles: candidate particles and empty circles: field particles. In case (a), students are supposed to attend only lectures of the teacher. In cases (b) and (c), students are also engaged in collaborative work forming homogeneous and heterogeneous groups of two individuals, respectively.

- The probability for a student to learn is proportional to the level of knowledge of the teacher, \bar{u} . The subsequent increase of knowledge is a uniform random variable given by a fraction, Δ , of what he does not know, $1 - u_*$. The probability for a student to unlearn is proportional to the teacher's inability which is measured by $1 - \bar{u}$. The subsequent decrease of knowledge is a uniform random variable given by a fraction, Δ , of what he knows, u_* . Accordingly:

$$\mathcal{C}_j^j(u_* \rightarrow u|u_*, u^* = \bar{u}) = \bar{u} U_{[u_*, u_* + (1-u_*)\Delta]} + (1 - \bar{u}) U_{[u_*(1-\Delta), u_*]}. \quad (15)$$

For the student–student interaction

- The probability for a student to learn is proportional to the product of his level of ignorance, $1 - u_*$, and the level of knowledge of the interaction partner, u^* . The subsequent increase of knowledge is a uniform random variable given by a fraction, Δ , of what he does not know, $1 - u_*$. The probability for a student to keep his level of knowledge, u_* , is proportional to the level of knowledge itself. The probability for a student to unlearn is proportional to the product of the levels of ignorance of the student, $1 - u_*$, and of the interaction partner, $1 - u^*$. The subsequent decrease of knowledge is a uniform random variable given by a fraction, Δ , of what he knows, u_* . Accordingly:

$$\begin{aligned} \mathcal{A}_{j,k}^j(u_* \rightarrow u|u_*, u^*) = & (1 - u_*) u^* U_{[u_*, u_* + (1-u_*)\Delta]} + u_* \delta(u - u_*) \\ & + (1 - u_*) (1 - u^*) U_{[u_*(1-\Delta), u_*]}, \end{aligned} \quad (16)$$

where $U_{[a,b]}$ is the uniform probability density on the interval $[a, b]$.

Remark 3 According to Eqs. (15) and (16), due to the interactions, the level of knowledge of a student can not only increase but also decrease, which is not unexpected. Indeed, inappropriate teaching material, lack of attention of the students, disordered discussions, misunderstandings, and so on may result in unlearning.

Model's parameters: The model includes the following two parameters:

- $0 \leq \bar{u} < 1$: Teacher's level of knowledge and
- $\Delta \geq 0$: Maximum increase or decrease of knowledge.

5.1.2 Numerical Results

The kinetic models defined by Eq. (10) with Eqs. (11)–(12) are solved by means of a Monte Carlo particle method. Accordingly, the distribution function of each functional subsystem is represented by a number of computational particles which interact by stochastic rules defined by the transition probability densities given by Eqs. (15) and (16). It is assumed that, initially, the level of knowledge of LA and HA students is distributed according to a Gaussian of mean 0.2, 0.4, respectively, and variance 0.001. Moreover, it has been set $\bar{u} = 0.9$ and $\Delta = 0.2$.

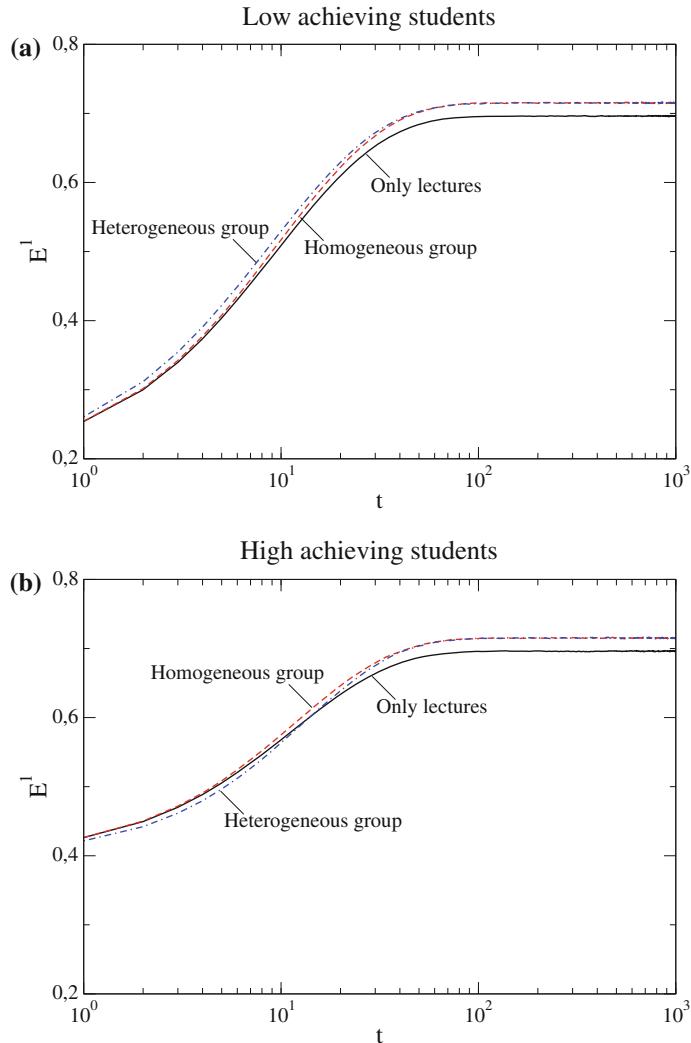


Fig. 4 Time evolution of the knowledge of (a) low-achieving students and (b) high-achieving students. The t -axis is in logarithmic scale. $\bar{u} = 0.9$, $\Delta = 0.2$.

Figure 4 shows the time evolution of the mean value of the students' knowledge. Three distinct time regimes can be distinguished. For the short time regime, $t \lesssim 10$, the knowledge increases rapidly. In the intermediate time regime, $10 \lesssim t \lesssim 100$, a progressive growth of the knowledge is observed. Finally, for the long time regime, $t \gtrsim 100$, the knowledge reaches a stationary maximum value. A direct inspection of the figure also shows that model is able to reproduce the well-established empirical evidence described in Reference [25]. More specifically, the achievement of the

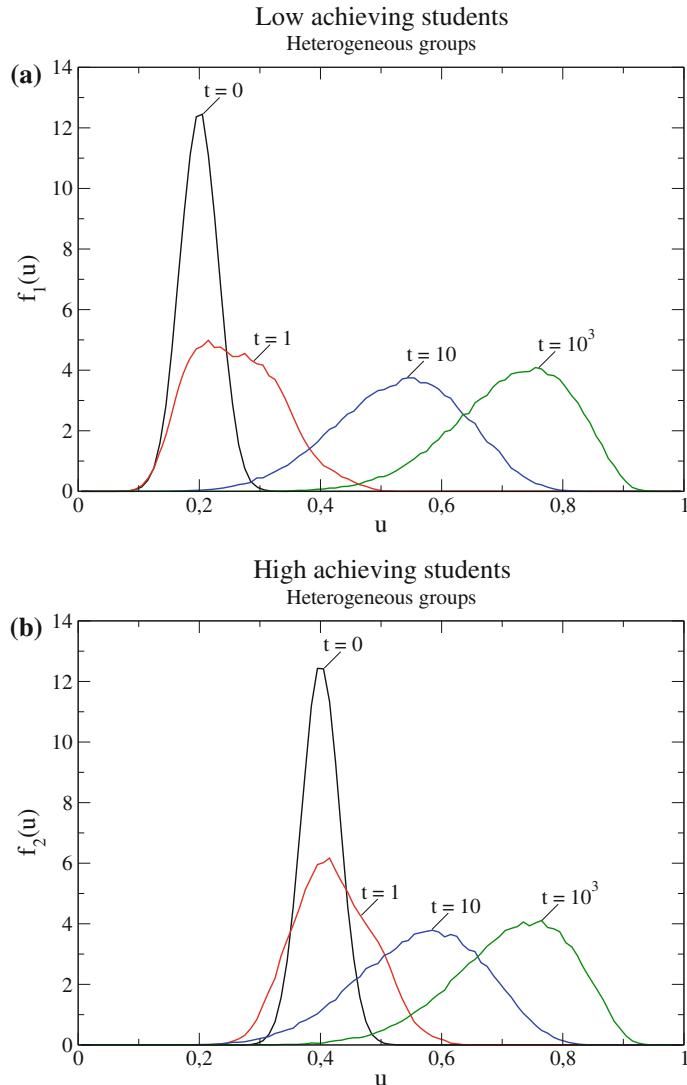


Fig. 5 Distribution function of (a) low-achieving students and (b) high-achieving students forming heterogeneous groups for $t = 0, 1, 10$ and 100 . $\bar{u} = 0.9$, $\Delta = 0.2$.

students involved in collaborative work is better than that of those attending only the lectures. Furthermore, the performance of LA students may be enhanced when they form groups with better students, although this improvement is obtained at the expense of the achievements of their colleagues.

In comparison with alternative approaches, the proposed kinetic theory modeling offers deeper insights into system dynamics by naturally providing the time evolution

of the density distribution functions of each group of students. Figure 5 shows the distribution functions of LA and HA students for the case (c) at the times $t = 0, 1, 10$ and 1000 . Independently on the initial achievement of the students, the distribution functions strongly modify. Afterward, their mean values increase, but the shapes of the distribution functions remain almost unchanged. A correspondence can be clearly identified between this dynamics and the three time regimes discussed above.

6 Modeling Social Systems

In recent years, a growing interest has arisen to study the collective emerging behaviors of social systems such as cooperation, cultural conflicts, and problems of social consensus.

Two complementary ways of modeling social systems can be roughly distinguished. The first approach involves the design and analysis of simplified mathematical models that do not try to mimic the real behavior of individuals but abstract the most important qualitative aspects so as to gain insight into the system dynamics. The rationale behind this approach is that in analogy with the “universality” concept in statistical physics, certain aspects of complex behavior are supposed to be independent on the specific dynamical details. The second approach is to design more comprehensive and realistic models, usually in the form of numerical simulations, which represent the interacting parts of a complex system, often down to minute details. The border between these two approaches is not sharply defined, and tools of each of them are always more often applied together.

The first approach has been mainly adopted to investigate social systems using methods of the statistical physics. Two research fields have been rapidly grown which are referred to as econophysics [82] and sociophysics [29, 43, 80]. Although the principles of both fields have much in common, econophysics focuses on the narrower subject of economic behavior, while sociophysics studies a broader range of social issues, including social networks, language evolution, population dynamics, epidemic spreading, terrorism, voting, coalition formation, and opinion dynamics.

A more realistic description of real-world social interactions can be build within game-theoretic models. Topics in economics studied using the methods of evolutionary game theory range from behavior in markets [76] to bargaining [83] to questions of public good provision and collective action [64]. Applications to problems of broader social science interest include residential segregation [84], cultural evolution [23], and the study of behavior in transportation [73].

As previously mentioned, in order to fully account for heterogeneity of living systems, agent-based modeling has been increasingly used. Since the very simple model on racial segregation proposed by Schelling [72], a plethora of models has been developed, which have taken into account several features such as bounded rationality, personal well-being, wealth and social status/education, and many others [6, 36].

Methods of classical kinetic theory have been also applied to model various social systems from the pioneer papers [19, 20]. Nowadays, the most important reference is given by the book by Pareschi and Toscani [68], which provides an exhaustive overview covering the whole path from numerical methods to mathematical tools and applications, such as opinion formation and wealth dynamics, as well as various aspects of the social and economical dynamics of our society.

Generalized kinetic theory and game theoretical tools are at the basis of the modeling approach to socioeconomic complex systems, which rests on the methods of the kinetic theory of active particles [1, 2]. In the next section, it is shown how such an approach can provide a realistic description of the criminality dynamics.

6.1 Case Study: *Criminality onset and Development*

The mathematical literature on the dynamics of criminality is constantly growing due to the impact on the well-being and security of citizens. Useful references are given by the following essays [17, 38, 39]. The interplay between wealth distribution and unethical behaviors is an object of growing interest of researchers, who operate in the field of social sciences [55], while it is becoming a new field of investigation of applied mathematicians [12, 35].

The case study presented in the following has been developed in [14]. It discusses the onset and the development of criminality in a society and, more specifically, how the criminality can be contrasted by acting, on the one side, upon the social state of citizens, such as wealth and culture, and, on the other side, upon the training of security forces.

The rest of the subsection is organized into two parts. The first part develops the mathematical formulation of the test case, while the second part briefly presents and discusses the numerical results.

6.1.1 Mathematical formulation

The model developed in [14] does not account for space dynamics, and hence, the system comprises only one node. Three functional subsystems are considered: citizens ($j = 1$), criminals ($j = 2$), and detectives ($j = 3$). The activity variables expressed by them are, respectively, wealth, criminal ability, and detective ability.

The mathematical structure is obtained by a specialization of Eqs. (5)–(7). In detail:

$$\partial_t f_j(t, u) = A_j[\mathbf{f}](t, u) + B_j[\mathbf{f}](t, u) \quad (17)$$

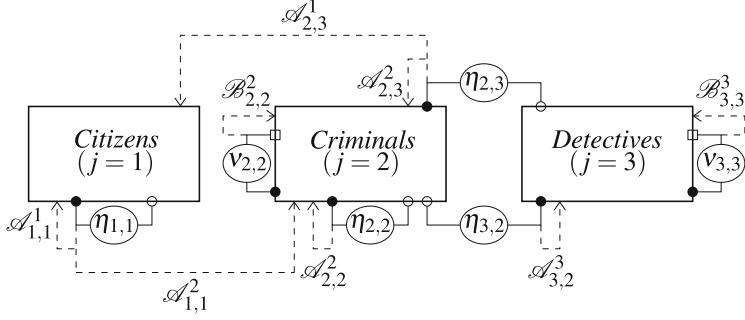


Fig. 6 Interaction rates, $\eta_{p,q}$ and $v_{p,q}$ and transition probability densities, $\mathcal{A}_{p,q}^j$ and $\mathcal{B}_{p,q}^j$. Solid circles: candidate particles; empty circles: field particles; empty squares: mean value of the activity within the functional subsystem.

where

$$\begin{aligned} A_j[\mathbf{f}](t, u) = & \sum_{q,k=1}^3 \int_{D_u \times D_u} \eta_{q,k}(u_*, u^*) \mathcal{A}_{q,k}^j(u_* \rightarrow u | u_*, u^*) f_q(t, u_*) f_k(t, u^*) du_* du^* \\ & - f_j(t, u) \sum_{k=1}^3 \int_{D_u} \eta_{j,k}(u, u^*) f_k(t, u^*) du^* \end{aligned} \quad (18)$$

and

$$\begin{aligned} B_j[\mathbf{f}](t, u) = & \int_{D_u} v_{j,j}(u_*, \mathbb{E}_j^1) \mathcal{B}_{j,j}^j(u_* \rightarrow u | u_*, \mathbb{E}_j^1) f_j(t, u_*) du_* \\ & - v_{j,j}(u, \mathbb{E}_j^1) f_j(t, u) \end{aligned} \quad (19)$$

The social structure of citizens is assumed to be fixed; namely, the time interval is supposed sufficiently short that the wealth distribution can be considered constant in time. Interaction rates and transition probability densities are summarized in Fig. 6 and are discussed in some more depth in the following. For more details, the reader is referred to [14].

Interaction Rates:

For citizens' interactions

- Interaction rate is higher for citizens with closer social states:

$$\eta_{11}(u_*, u^*) = \eta^0 (1 - |u_* - u^*|). \quad (20)$$

For criminals' interactions

- Interaction rate between criminals is higher for experienced criminals:

$$\eta_{22}(u_*, u^*) = \eta^0(u_* + u^*). \quad (21)$$

- Interaction rate between criminals and detectives is higher for inexperienced criminals and experienced detectives:

$$\eta_{23}(u_*, u^*) = \eta^0((1 - u_*) + u^*). \quad (22)$$

- Interaction rate between criminals and the lawbreakers' environment is higher for criminals whose ability is lower than the mean ability of criminals:

$$\nu_{2,2}(u_*, \mathbb{E}_2^1) = \begin{cases} 2v^0|u_* - \mathbb{E}_2^1| & u_* < \mathbb{E}_2^1 \\ 0 & u_* > \mathbb{E}_2^1 \end{cases}. \quad (23)$$

For detectives' interactions

- Interaction rate between detectives and criminals is higher for experienced detectives and inexperienced criminals:

$$\eta_{32}(u_*, u^*) = \eta^0(u_* + (1 - u^*)). \quad (24)$$

- Interaction rate between detectives and the security forces' environment is higher for detectives whose ability is lower than the mean ability of detectives:

$$\nu_{3,3}(u_*, \mathbb{E}_3^1) = \begin{cases} 2v^0|u_* - \mathbb{E}_3^1| & u_* < \mathbb{E}_3^1 \\ 0 & u_* > \mathbb{E}_3^1 \end{cases}. \quad (25)$$

Transition Probability Densities:

For citizens' interactions

- Citizens are susceptible to become criminals, motivated by their wealth state. More in detail, a candidate citizen with activity u_* interacting with a richer one with activity $u^* > u_*$ can become a criminal, mutating into functional subsystem $j = 2$ with a very low criminal ability $u = \varepsilon \approx 0$. In particular, it is assumed that the transition probability increases with decreasing wealth:

$$\mathcal{A}_{11}^2(u_* \rightarrow u | u_*, u^*) = \frac{1}{\varepsilon} \alpha_1 (1 - u_*) u^* \chi_{[0, \varepsilon)}(u), \quad (26)$$

where $\chi_{[0, \varepsilon)}$ denotes the indicator function for the interval $[0, \varepsilon)$.

For criminals' interactions

- Less experience criminals mimic the more experienced ones, but also more experienced criminals, due to interactions, may increase their ability:

$$\mathcal{A}_{22}^2(u_* \rightarrow u | u_*, u^*) = \delta(u - (u_* + \beta_1(1 - u_*)u^*)). \quad (27)$$

- Criminals chased by detectives are constrained to step back, decreasing their activity value as the price to be paid for being caught. Furthermore, criminals are induced to return to the state of normal citizens with probability which increases with decreasing the values of their level of criminality and increasing the values of detective skills:

$$\begin{cases} \mathcal{A}_{23}^1(u_* \rightarrow u|u_*, u^*) = \frac{1}{\varepsilon} \alpha_2 (1 - u_*) u^* \chi_{[0, \varepsilon)}(u), \\ \mathcal{A}_{23}^2(u_* \rightarrow u|u_*, u^*) = (1 - \alpha_2(1 - u_*) u^*) \delta(u - (u_* - \gamma u^* u_*)). \end{cases} \quad (28)$$

- Criminals show a trend toward the mean ability of the lawbreakers' environment:

$$\mathcal{B}_2(u_* \rightarrow u|u_*, \mathbb{E}_2^1) = \delta(u - (\beta_1 u_* + (1 - \beta_1) \mathbb{E}_2^1)). \quad (29)$$

For detectives' interactions

- Detectives gain experience by chasing criminals and increase their ability:

$$\mathcal{A}_{32}^3(u_* \rightarrow u|u_*, u^*) = \delta(u - (u_* + \gamma u^*(1 - u_*))). \quad (30)$$

- Detectives show a trend toward the mean ability of the security forces' environment:

$$\mathcal{B}_3(u_* \rightarrow u|u_*, \mathbb{E}_3^1) = \delta(u - (\beta_2 u_* + (1 - \beta_2) \mathbb{E}_3^1)). \quad (31)$$

Model's parameters: The model includes the following seven parameters:

- $\eta^0 > 0$: Constant factor for microscopic interactions;
- $v^0 > 0$: Constant factor for microscopic–macroscopic interactions;
- $0 \leq \alpha_1 < 1$: Susceptibility of citizens to become criminals;
- $0 \leq \alpha_2 < 1$: Susceptibility of criminals to get back the state of normal citizens;
- $0 \leq \beta_1 < 1$: Learning dynamics among criminals;
- $0 \leq \beta_2 < 1$: Learning dynamics among detectives;
- $0 \leq \gamma < 1$: Motivation/efficacy of security forces to catch criminals.

6.1.2 Numerical results

The kinetic models defined by Eqs. (17)–(19) are solved by means of a deterministic method of solution. The distribution function of each functional subsystem is approximated at a number of collocation points over the activity variable. This transforms the original integro-differential system into a coupled system of ordinary differential equations which is then solved by classical numerical method.

Simulations have been carried out corresponding to fixed values of the mean wealth of citizens. More specifically, two values have been selected, low ($\mathbb{E}_1^1 = 0.2$)

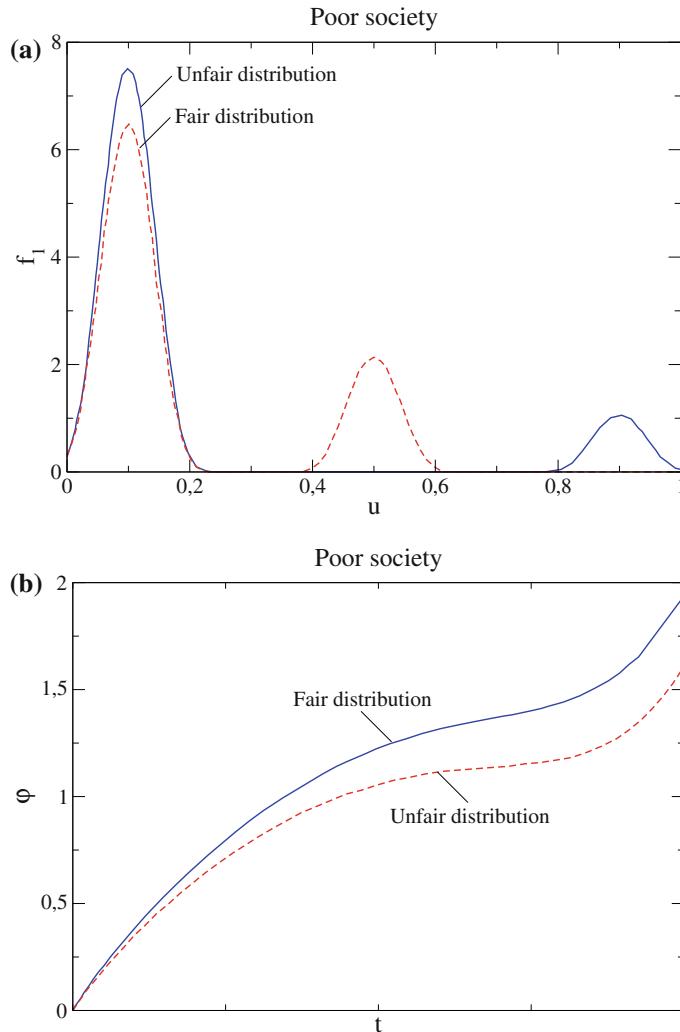


Fig. 7 (a) Initial wealth distributions for a poor society with $E_1 = 0.2$. (b) Relative change of the population of criminals. $\alpha_1 = 0.0001$, $\alpha_2 = 0.15$, $\beta_1 = 0.1$, $\beta_2 = 0.9$, $\gamma = 0.15$.

and high ($E_1^1 = 0.6$), while two different shapes have been considered for each case corresponding to higher and lower concentrations of wealth in the middle class, as depicted in Figs. 7(a) and 8(a). As initial condition, it has been assumed that $n_2(0)/n_1(0) = 0.05$ and $n_3(0)/n_1(0) = 0.005$, corresponding to a society where the initial number of criminals is 5% of the number of citizens and the number of detectives is 10% of the number of criminals. The following quantity has been defined measure of the relative percentage change in the population of criminals

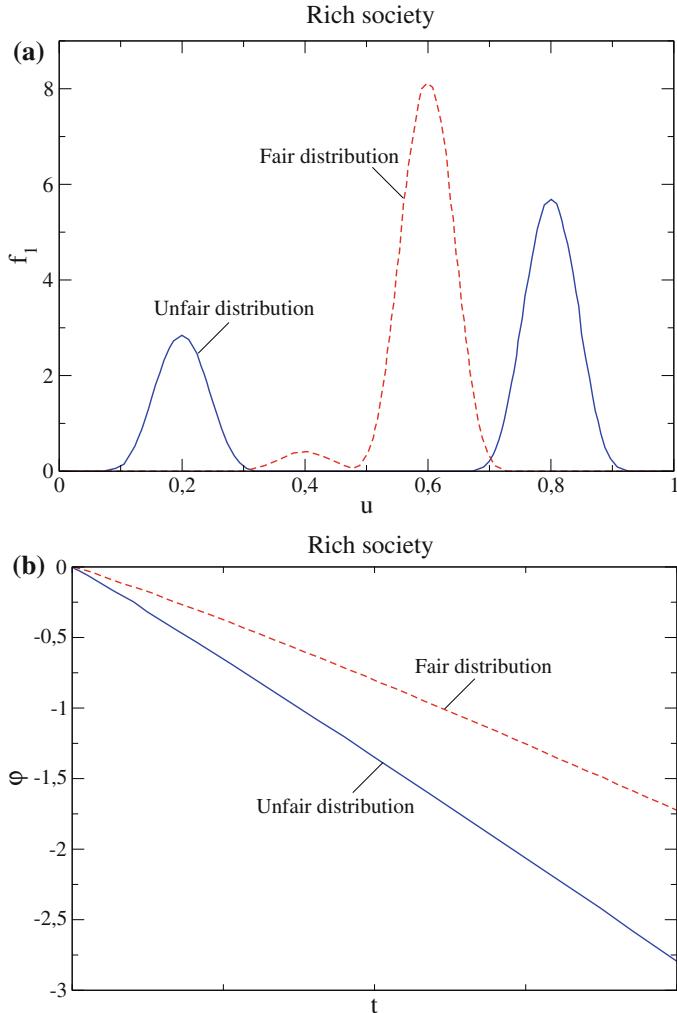


Fig. 8 (a) Initial wealth distributions for a rich society with $\mathbb{E}_1 = 0.6$. (b) Relative change of the population of criminals. $\alpha_1 = 0.0001$, $\alpha_2 = 0.15$, $\beta_1 = 0.1$, $\beta_2 = 0.9$, $\gamma = 0.15$.

$$\varphi(t) = \frac{1}{n_2(0)}(n_2(t) - n_2(0)) \times 10^2. \quad (32)$$

Figures 7(b) and 8(b) report the evolution of φ . We can observe that a poorer society produces a growth in the number of criminals, that is still more accentuated for unequal wealth distributions. The model is capable of producing the opposite behavior for a richer society, giving a reduction in the number of criminals for the same choice of parameters.

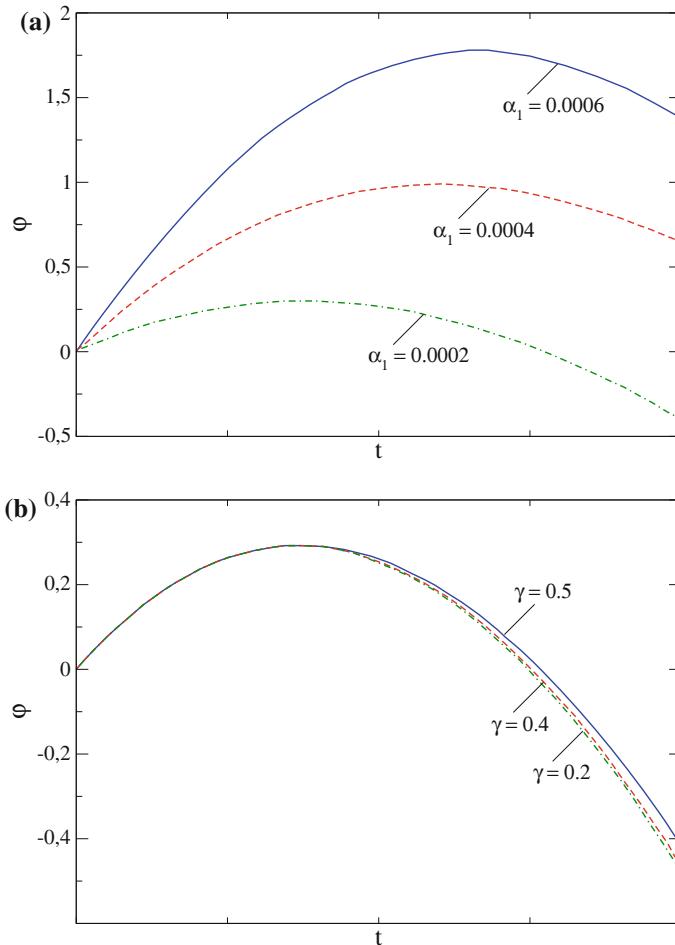


Fig. 9 (a) Relative change of the population of criminals for different values of citizens susceptibility to become criminals. (b) Relative change of the population of criminals for different values of motivation/efficacy of security forces to catch criminals. $\beta_1 = 0.1$, $\beta_2 = 0.9$, $\gamma = 0.5$.

An additional interesting topic consists in assessing whether it is better to improve the well-being of the society or to strengthen the ability of security forces. The initial distribution of detectives is assumed to be a Gaussian-type function with mean value 0.5. Figure 9(a) shows the time dynamics of φ for different values of α_1 and $\alpha_2 = 0.05$, while Figure 9(b) shows the same dynamics for different values of α_2 and $\alpha_1 = 0.0002$. These results confirm the empirical evidence that an effective action to fight crime consists in pursuing actions that contribute to reduce α_1 (education, employment, etc) and to improve the quality of citizens [67].

7 Looking Ahead to Research Perspectives

This chapter has shown how methods of the mathematical kinetic theory and theoretical tools of game theory can be developed to model real systems in the field of the so-called soft sciences. The first part of the chapter has reviewed and critically analyzed mathematical tools to pursue this challenging objective, while the second part has presented two specific applications, namely collective learning dynamics and a complex interaction involving citizens, criminals, and police forces.

The chapter can now be closed by proposing to the readers' attention some research perspectives which appear, according to the authors' bias, worth of future studies.

Nonlinear interactions: Nonlinearity of interactions at the microscale means that the output of interactions could depend not only on the microscopic state of the interacting entities, but also on the probability distribution functions of the functional subsystems to which they belong. This topic has been introduced, in the field of social sciences, in [12] and [35]. Both papers analyze the interplay of different dynamics, respectively, wealth distribution versus support or opposition to governments, and selfishness in wealth dynamics as a source of overall wealth dissipation in nations. An interesting reference to the social aspects of selfishness and wealth distribution is given by [70], based on previous studies in the field [54, 55].

Nonlinearity at the microscopic scale introduces additional nonlinearity in the mathematical structure of models and, as shown, by the aforementioned applications, an important influence on the predictive ability of models. However, the present knowledge is limited to heuristic assumptions linked to the said papers, while a systematic analysis appears worth to be developed.

Mutations and selection: Post-Darwinian dynamics consisting in mutations followed by selection plays an important role in biology, specifically in immune competition [16, 33]. An analogous dynamics appears in socioeconomic systems, where new groups can be generated, for instance by the aggregation of different groups which subsequently might either expand or disappear in a competition mediated by the external environment. These concepts have been introduced in various recent papers, for instance [32] in a behavioral theory of urbanism.

The need of further studies in this topic has already been motivated in [3], but not yet developed in a systematic approach. It can be argued that mutations can be modeled by allowing the transition probability densities to have a stochastic output across functional subsystems and even into new subsystems. Selection can be related to the fitness of the newborn functional systems with the external environment. Therefore, some of them survive and even expand, while others show a trend to extinction.

These two topics, as already mentioned, do not claim to provide an exhaustive panorama of possible future research activity, but represent a challenging research perspective which already involves the authors of this chapter. More in general, it might attract applied mathematicians involved in the challenging objective of modeling living systems composed of many entities.

References

1. Ajmone Marsan, G.: On the modelling and simulation of the competition for a secession under media influence by active particles methods and functional subsystems representation **57**(5), 710–728 (2009)
2. Ajmone Marsan, G., Bellomo, N., Egidi, N.: Towards a mathematical theory of complex socio-economical systems by functional subsystems representation. *Kinetic Rel. Models* **1**(2), 249–278 (2008)
3. Ajmone Marsan, G., Bellomo, N., Gibelli, L.: Towards a systems approach to behavioral social dynamics. *Math. Mod. Meth. Appl. Sci.* **26**(6), 1051–1093 (2016)
4. Allen, B., Nowak, M.A.: Games on graphs. *EMS Surv. Math. Sci.* **1**(1), 113–151 (2014)
5. Anderson, P.W.: More is different. *Science* **177**(4047), 393–396 (1972)
6. Axelrod, R.: *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton University Press, Princeton (1997)
7. Ballerini, M., Cabibbo, N., Candelier, R., Cisbani, E., Giardina, I., Lecomte, V., Orlandi, A., Parisi, G., Procaccini, A., Viale, M., Zdravkovic, V.: Interaction ruling animal collective behavior depends on topological rather than metric distance: evidence from a field study. *Proc. Natl. Acad. Sci. USA* **105**(4), 1232–1237 (2008)
8. Barabasi, A.L.: *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science and Everyday Life*. Plume Editors (2002)
9. Barbante, P., Frezzotti, A., Gibelli, L.: A kinetic theory description of liquid menisci at the microscale. *Kinetic Rel. Models* **8**(2), 235–254 (2015)
10. Bellomo, N.: *Modeling Complex Living Systems - A Kinetic Theory and Stochastic Game Approach*. Birkhäuser, Boston, New York (2008)
11. Bellomo, N., Soler, J.: On the mathematical theory of the dynamics of swarms viewed as complex systems. *Math. Mod. Meth. Appl. Sci.* **22**, Paper n.1140,006 (2012)
12. Bellomo, N., Herrero, M.A., Tosin, A.: On the dynamics of social conflicts: looking for the black swan. *Kinetic Rel. Models* **6**, 459–479 (2013)
13. Bellomo, N., Knopoff, D., Soler, J.: On the difficult interplay between life, “complexity”, and mathematical sciences. *Math. Mod. Meth. Appl. Sci.* **23**, 1861–1913 (2013)
14. Bellomo, N., Colasuonno, F., Knopoff, D., Soler, J.: From a systems theory of sociology to modeling the onset and evolution of criminality. *Netw. Heterog. Media* **10**(3), 421–441 (2015)
15. Bellouquid, A., De Angelis, E., Ferme, L.: Towards the modeling of vehicular traffic as a complex system: A kinetic theory approach. *Math. Mod. Meth. Appl. Sci.* **22**(5), Paper n.1140,003 (2012)
16. Bellouquid, A., De Angelis, E., Knopoff, D.: From the modeling of the immune hallmarks of cancer to a black swan in biology. *Math. Mod. Meth. Appl. Sci.* **23**, 949–978 (2013)
17. Berenji, B., Chou, T., D’Orsogna, M.R.: Recidivism and rehabilitation of criminal offenders: A carrot and stick evolutionary game. *PLOS ONE* **9**, Paper n.885,531 (2014)
18. Berliner, D.C., Calfee, R.C. (eds.): *Handbook of Educational Psychology*. MacMillan, New York (1996)
19. Bertotti, M.L., Delitala, M.: From discrete kinetic and stochastic game theory to modelling complex systems in applied sciences. *Math. Mod. Meth. Appl. Sci.* **14**(7), 1061–1084 (2004)
20. Bertotti, M.L., Delitala, M.: Conservation laws and asymptotic behavior of a model of social dynamics. *Nonlinear Anal. Real World Appl.* **9**(1), 183–196 (2008)
21. Binder, K. (ed.): *Applications of the Monte Carlo Method in Statistical Physics*. Springer-Verlag, Heidelberg (1987)
22. Bird, G.A.: *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Oxford University Press, Oxford (UK) (1994)
23. Bisin, A., Verdier, T.: The economics of cultural transmission and the dynamics of preferences. *J. Econ. Theory* **97**, 298–319 (2001)
24. Bonabeau, E., Dorigo, M., Theraulaz, G.: *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, Oxford (1999)

25. Bordogna, C.M., Albano, E.V.: Theoretical description of teaching-learning processes: A multidisciplinary approach. *Phys. Rev. Lett.* **87**(11), 118,701 (2001)
26. Burini, D., De Lillo, S., Gibelli, L.: Collective learning dynamics modeling based on the kinetic theory of active particles. *Phys. Life Rev.* **16**, 123–139 (2016)
27. Bush, R.R., Mosteller, F.: Stochastic models for learning. John Wiley & Sons, Inc., Oxford (1955)
28. Camerer, C.F.: Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton (2003)
29. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009)
30. Couzin, I.D.: Collective minds. *Nature* **445**, 715 (2007)
31. Cucker, F., Smale, S.: On the mathematical foundations of learning. *Bull. Amer. Mathe. Soc.* **39**(1), 1–49 (2001)
32. D'Acci, L.: Mathematize *urbes* by humanizing them. Cities as isobenefit landscapes: psycho-economical distances and personal isobenefit lines. *Landscape Urban Plan.* **139**, 63–81 (2015)
33. De Angelis, E.: On the mathematical theory of post-Darwinian mutations, selection, and evolution. *Math. Mod. Meth. Appl. Sci.* **24**(13), 2723–2742 (2014)
34. Dimarco, G., Pareschi, L.: Numerical methods for kinetic equations. *Acta Num.* **23**(12), 369–520 (2014)
35. Dolfin, M., Lachowicz, M.: Modeling altruism and selfishness in welfare dynamics: the role of nonlinear interactions. *Math. Mod. Meth. Appl. Sci.* **24**, 2361–2381 (2014)
36. Epstein, J.M., Axtell, R.: Growing artificial societies: social science from the bottom up. The MIT Press, Boston (1996)
37. Estes, W.K., Suppes, P.: Foundations of Statistical Learning Theory, II. The Stimulus Sampling Model. Tech. Rep. 26, Stanford University, Institute for Mathematical Studies in the Social Sciences, Stanford University (1959)
38. Fajnzylber, P., Lederman, D., Loayza, N.: Inequality and violent crime. *J. Law. Econ.* **45**(1), 1–40 (2002)
39. Felson, M.: What every mathematician should know about modelling crime. *Eur. J. Appl. Math.* **21**, 275–281 (2010)
40. Frezzotti, A., Ghiraldi, G.P., Gibelli, L.: Solving model kinetic equations on GPUs. *Comput. Fluids* **50**, 136–146 (2011)
41. Frezzotti, A., Ghiraldi, G.P., Gibelli, L.: Solving the Boltzmann equation on GPUs. *Comput. Phys. Commun.* **182**, 2445–2453 (2011)
42. Frezzotti, A., Ghiraldi, G.P., Gibelli, L., Bonucci, A.: DSMC simulation of rarefied gas mixtures flows driven by arrays of absorbing plates. *Vacuum* **103**, 57–67 (2014)
43. Galam, S.: Sociophysics: A Physicists Modeling of Psycho-political Phenomena. Springer, New York (2012)
44. Galam, S., Moscovici, S.: Towards a theory of collective phenomena: Consensus and attitude changes in groups. *Eur. J. Soc. Psy.* **21**, 49–74 (1991)
45. Gintis, H.: The bounds of reason: Game Theory and the unification of the behavioral sciences. Princeton University Press, Princeton and Oxford (2009)
46. Gintis, H.: Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction. Princeton University Press, Princeton and Oxford (2009)
47. Harsanyi, J.C.: Games with incomplete information played by “bayesian” players, i-iii part i. the basic model. *Manage. Sci.* **14**(3), 159–182 (1967)
48. Helbing, D.: Stochastic and Boltzmann-like models for behavioral changes, and their relation to game theory. *Physica A* **193**(2), 241–258 (1993)
49. Helbing, D.: Quantitative sociodynamics: Stochastic methods and models of social interaction processes. Springer Verlag (2010)
50. Helbing, D., Yu, W.: The outbreak of cooperation among success-driven individuals under noisy conditions. *Proc. Nat. Acad. Sci.* **106**(10), 3680–3685 (2009)
51. Hofbauer, J., Sigmund, K.: Evolutionary game dynamics. *Bull. Amer. Mathe. Soc.* **40**(4), 479–519 (2003)

52. Knopoff, D.: On the modeling of migration phenomena on small networks. *Math. Mod. Meth. Appl. Sci.* **23**(3), 541–563 (2013)
53. Knopoff, D.: On a mathematical theory of complex systems on networks with application to opinion formation. *Math. Mod. Meth. Appl. Sci.* **24**(2), 405–426 (2014)
54. Kraus, M.W., Piff, P.K., Keltner, D.: Social class, sense of control, and social explanation. *J. Pers. Soc. Psychol.* **97**(6), 992–1004 (2009)
55. Kraus, M.W., Piff, P.K., Mendoza-Denton, R., Rheinschmidt, M.L., Keltner, D.: Social class, solipsism, and contextualism: How the rich are different from the poor. *Psychol. Rev.* **119**(3), 546–572 (2012)
56. Lasry, J.M., Lions, P.L.: Jeux à champ moyen. i le cas stationnaire. *CR. Acad. Sci. I-Math.* **343**(9), 619–625 (2006)
57. Lasry, J.M., Lions, P.L.: Jeux à champ moyen. ii horizon fini et contrôle optimal. *CR. Acad. Sci. I-Math.* **343**(10), 679–684 (2006)
58. Latané, B.: The psychology of social impact. *Am. Psychol.* **36**(4), 343–356 (1981)
59. Lave, J., Wenger, E.: *Situated Learning: Legitimate Peripheral Partecipation*. Cambridge University Press, Cambridge (1998)
60. May, R.M.: Uses and abuses of mathematics in biology. *Science* **303**, 790–793 (2004)
61. Mayr, E.: *What Evolution Is*. Basic Books, New York (2001)
62. Morgenstern, O., Von Neumann, J.: *Theory of Games and Economic Behavior*. Princeton University Press (1953)
63. Motsch, S., Tadmor, E.: Heterophilious dynamics enhances consensus. *Siam Rev.* **56**(4), 577–621 (2014)
64. Myatt, D.P., Wallace, C.: An evolutionary analysis of the volunteer's dilemma. *Game Econ. Behav.* **62**, 67–76 (2008)
65. Nash, J.: Non-cooperative games. *Ann. Math.* **54**(2), 286–295 (1951)
66. Nowak, M.A.: *Evolutionary Dynamics. Exploring the Equations of Life*. Harvard University Press, Cambridge (MA) (2006)
67. Ormerod, P.: Crime: Economic incentives and social networks. *IEA Hobart Paper* **151**(4), 1–54 (2005)
68. Pareschi, L., Toscani, G.: *Interacting Multiagent Systems - Kinetic equations and Monte Carlo methods*. Oxford University Press, Oxford (UK) (2013)
69. Piaget, J.: *The Child's Conception of the World*. Harcourt, Brace, New York (1929)
70. Piff, P.K., Stancato, D.M., Côté, S., Mendoza-Denton, R., Keltner, D.: Higher social class predicts increased unethical behavior. *Proc. Natl. Acad. Sci. USA* **109**(11), 4086–4091 (2014)
71. Rufus, I.: *Differential games: A mathematical theory with applications to warfare and pursuit, control and optimization*. John Wiley and Sons, New York (1965)
72. Schelling, T.C.: Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186 (1971)
73. Smith, M.J.: The stability of a dynamic model of traffic assignment - an application of a method of Lyapunov. *Transport. Sci.* **18**(3), 245–252 (1984)
74. Szabó, G., Fáth, G.: Evolutionary games on graphs. *Phys. Rep.* **446**, 97–216 (2007)
75. Taleb, N.N.: *The Black Swan: The Impact of the Highly Improbable*. Random House, New York City (2007)
76. Vega-Redondo, F.: *Evolution, Games, and Economic Behaviour*. Oxford University Press, Oxford (1996)
77. Vygotsky, L.S.: *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, Cambridge, Massachusetts London, England (1978)
78. Wagner, W.: A convergence proof of bird's direct simulation Monte-Carlo method for the Boltzmann equation. *J. Stat. Phys.* **66**, 1011–1044 (1992)
79. Webb, G.F.: *Theory of Nonlinear Age-dependent Population Dynamics*. Marcel Dekker, New York (1985)
80. Weidlich, W., Haag, G.: *Concepts and Models of a Quantitative Sociology The Dynamics of Interacting Populations*. Springer-Verlag, Berlin (1983)
81. Wolpert, D.H.: Information theory - the bridge connecting bounded rational game theory and statistical physics. In: D. Braha, A.A. Minai, Y. Bar-Yam (eds.) *Complex Engineered Systems, Understanding complex systems*, pp. 262–290. Springer, Princeton (2006)

82. Yakovenko, V.M., Barkley Rosser, J.: Statistical physics of social dynamics. *Rev. Mod. Phys.* **4**, 1703–1725 (2009)
83. Young, H.P.: An evolutionary model of bargaining. *J. Econ. Theory* **59**, 145–168 (1993)
84. Zhang, J.: A dynamic model of residential segregation. *J. Math. Sociol.* **28**, 147–170 (2004)

A Review on Attractive–Repulsive Hydrodynamics for Consensus in Collective Behavior

José A. Carrillo, Young-Pil Choi and Sergio P. Perez

Abstract This survey summarizes and illustrates the main qualitative properties of hydrodynamics models for collective behavior. These models include a velocity consensus term together with attractive–repulsive potentials leading to non-trivial flock profiles. The connection between the underlying particle systems and the swarming hydrodynamic equations is performed through kinetic theory modeling arguments. We focus on Lagrangian schemes for the hydrodynamic systems showing the different qualitative behaviors of the systems and its capability of keeping properties of the original particle models. We illustrate the known results concerning large-time profiles and blowup in finite time of the hydrodynamic systems to validate the numerical scheme. We finally explore the unknown situations making use of the numerical scheme showcasing a number of conjectures based on the numerical results.

1 Introduction

Modeling the collective behavior of a large number of interacting individuals is a very challenging problem in animal behavior, pedestrian flow, cell adhesion and chemotaxis problems, and many other biological applications—see for instance [15, 26, 27, 53, 88, 89] and the literature therein. Most of the literature is based on Individual-based Models (IBMs) which are particle descriptions from a kinetic

J.A. Carrillo (✉)

Department of Mathematics, Imperial College London, South Kensington SW7 2AZ, UK
e-mail: carrillo@imperial.ac.uk

Y.-P. Choi

Fakultät für Mathematik, Technische Universität München, Boltzmannstraße 3,
85748 Garching bei München, Germany
e-mail: ychoi@ma.tum.de

S.P. Perez

ETSIAE, Technical University of Madrid, Pza. de Cardenal Cisneros, 3,
28040 Madrid, Spain
e-mail: sergio.perez.perez@alumnos.upm.es

modeling perspective. These particle systems typically include three basic effects: attraction, repulsion, and alignment or re-orientation of the individuals, called the first principles of swarming. The way in which these three effects are taken into account has given rise to a large number of different and interesting models for collective behavior. These basic 3-zone models were introduced by theoretical biologists [5, 72] for fisheries control as well as computer scientists [91] in order to mimic animal behavior in animation movies. These models have evolved toward more complete descriptions involving particular species interactions and adapted to particular animals such as birds [71], fish [14, 70, 73], ducks [81, 82], and insects [16] for instance.

These basic particle descriptions can be coarsened to macroscopic descriptions when the number of individuals is large leading to nonlocal macroscopic models both at the level of the mass density [83, 84] and hydrodynamic descriptions [38, 52]. This connection to continuum models is better done by passing to the intermediate description provided by kinetic modeling. The kinetic theory approach via mean-field limits of interacting particle systems has offered a mathematical underpinning to derive kinetic equations in a rigorous manner from particle descriptions. The connection toward macroscopic equations is done either via closure assumptions or via moment approximations [38, 52] and large friction limits [77]. One of the most famous particle models was introduced by Vicsek and his collaborators [99] showing a phase transition behavior that has also been studied through kinetic modeling and self-organized hydrodynamics [56–58]. We refer to [41, 75] and the references therein for a good account of the different levels of description and the state of the art of these models in the applied mathematical community.

We will focus in this review on two velocity consensus models [54, 55, 85, 86] that lead to asymptotic convergence for the large time toward a fixed velocity under certain conditions, a phenomenon that is called asymptotic flocking. These models have been studied extensively in the last years due to their apparent simplicity in formulating the possibility of consensus in velocity. These models are connected to the Vicsek model in which all particles travel to a fixed speed by large friction limits [24]. They also present a phase transition in terms of noise as the original Vicsek model [13]. In this survey, we concentrate in the basic properties of the hydrodynamic models incorporating also the effects of attraction and repulsion through an interaction potential.

In Section 2, we give a brief account of the particle descriptions making particular emphasis to the consensus in velocity models with interaction potentials and their flock solutions. Section 3 is first devoted to explain the link between these particle models and hydrodynamic descriptions via kinetic modeling. We propose a Lagrangian approach to solve the one-dimensional hydrodynamic descriptions. We numerically explore different qualitative aspects of the hydrodynamic models such as critical thresholds [34] and their sharpness for consensus models with and without interaction potentials. We also analyze the effect of the singularity of the potential in the longtime asymptotics of global solutions.

2 Microscopic Descriptions: Discrete Models

In this section, we review some of the basic individual-based attractive–repulsive models containing an additional velocity alignment force. The social interaction between individuals of the swarm is modeled by an effective interaction potential encapsulating the short-range repulsion and the long-range attraction forces at the particle level as discussed in the introduction. On top, we will also consider cases in which there is a tendency of behaving similarly to other individuals of the group, this mimicking behavior can be modeled in many different ways. One of the simplest manners of incorporating this gregarious behavior is to assume that each individual averages its relative velocity vector with nearby individuals according to some weights that we call the communication function. All the modeling in these simple descriptions are reduced to find biologically reasonable potentials and communication functions for the particular application or adapted to a particular species. Many authors have studied what are the most probable interaction regions for different animals, see [70, 73, 82] and the references therein. We will showcase some of the different behaviors in these models by choosing toy example for potentials and communication functions. Although not too biologically reasonable, these choices give us generic behaviors for these models.

For the velocity alignment force, we use two different types of forces proposed by Cucker and Smale [54, 55] and Motsch and Tadmor [85]. More precisely, let (x_i, v_i) be the position and velocity of i -th individual. Then our main system reads as

$$\begin{cases} \frac{dx_i}{dt} = v_i, & i = 1, \dots, N, \quad t > 0, \\ \frac{dv_i}{dt} = \frac{1}{S_i(x)} \sum_{j=1}^N \psi(x_i - x_j) (v_j - v_i) - \frac{1}{N} \sum_{j \neq i} \nabla K(x_i - x_j). \end{cases} \quad (1)$$

The first term on the right-hand side of (1)₂ represents a nonlocal velocity alignment force, where ψ is the communication function. The second term on the right-hand side of (1)₂ serves as attractive–repulsive forces through the interaction potential $K(x)$. Typical assumptions on $K(x)$ are radially symmetric and smooth outside the origin possibly decaying to zero for large distances, one particular example widely used in the literature is the Morse potential, see [3, 43, 47, 60] for more details. Here the scaling function $S_i(t)$ and the communication function ψ are given by

$$S_i(x) := \begin{cases} N & \text{for the Cucker-Smale model,} \\ \sum_{k=1}^N \psi(x_i - x_k) & \text{for the Motsch-Tadmor model,} \end{cases} \quad (2)$$

and

$$\psi(x) = \frac{1}{(1 + |x|^2)^{\beta/2}}, \quad \beta \geq 0,$$

respectively. These scalings are related to the mean-field limit for the system of the N interacting particles. Assuming that the effect of each individual on another one via the social force decays as $1/N$ is intuitive; if we want to obtain some non-trivial limit as $N \rightarrow \infty$, we should keep the total kinetic and potential energy and velocity of each individual to be of order 1 in that limit. More discussions about the mean-field limit can be found in [2, 22, 23, 25, 28–31, 41, 59, 63, 64, 68, 93, 94].

In [55], Cucker and Smale introduced the Newton-type particle system (1) for flocking phenomena. The local averaging of relative velocities is weighted by the communication function ψ in such a way that closer individuals have stronger influence than further ones. Note that the velocity alignment force of the Cucker–Smale (in short CS) model is scaled with the total mass. Later, Motsch and Tadmor proposed in [85] a new model for self-organized dynamics. They pointed out that the CS model is inadequate for far-from-equilibrium scenarios since the communication function is normalized by the total number of agents N . By taking into account the velocity alignment force normalized with a local average density, the Motsch–Tadmor (in short MT) model takes into account not only the relative distance between agents, but also their relative weights compared to the CS model. Note that the MT model does not have the symmetry property due to the normalization.

2.1 Velocity Alignment Models Without Interaction Forces

We begin our discussion with the case in which the individuals are only interacting through the velocity alignment force as

$$\begin{cases} \frac{dx_i}{dt} = v_i, & i = 1, \dots, N, \quad t > 0, \\ \frac{dv_i}{dt} = \frac{1}{S_i(x)} \sum_{j=1}^N \psi(x_i - x_j) (v_j - v_i), & \psi(x) = \frac{1}{(1 + |x|^2)^{\beta/2}}, \quad \beta \geq 0, \end{cases} \quad (3)$$

with the initial data

$$(x_i(0), v_i(0)) =: (x_{i0}, v_{i0}), \quad i = 1, \dots, N. \quad (4)$$

Here the scaling function $S_i(x)$ is given in (2). We notice that the standard Cauchy–Lipschitz theory yields the existence and uniqueness of global-in-time smooth solutions to the system (3) with $S_i(x) \equiv N$ since the communication function ψ is bounded and globally Lipschitz. For the MT model, we can also show that the communication function ψ is bounded from below for any time $T < \infty$, and this again enables us to apply the Cauchy–Lipschitz theory to the MT model to have the existence and uniqueness of solutions. Let us first remind the main analytical results concerning the flocking behavior for the system (3). Then, we present several numerical

results to illustrate the analytical ones. We also compare the time behavior of solutions to the CS and MT systems, i.e., (3) with $S_i(x) \equiv N$ and $S_i(x) = \sum_{k=1}^N \psi(x_i - x_k)$.

For the large-time behavior of solutions, we first introduce the definition of universal asymptotic flocking for the system (3).

Definition 1 Let $(x_i, v_i)_{i=1}^N$ be a given solution of the particle system (3)–(4). Then the $(x_i, v_i)_{i=1}^N$ leads to asymptotic flocking if and only if it satisfies the following two conditions:

$$\lim_{t \rightarrow \infty} \max_{1 \leq i, j \leq N} |v_i(t) - v_j(t)| = 0 \quad \text{and} \quad \sup_{0 \leq t < \infty} |x_i(t) - x_j(t)| < \infty.$$

We then define diameters in position and velocity phase spaces as follows:

$$R^x(t) := \max_{1 \leq i, j \leq N} |x_i(t) - x_j(t)|, \quad R^v(t) := \max_{1 \leq i, j \leq N} |v_i(t) - v_j(t)|. \quad (5)$$

For the system (3), rigorous estimates showing the emergence of flocking depending on the decay rate of the communication function are provided in [54, 55]. Later, the flocking estimates are refined in [40, 65, 66, 85, 96]. Flocking models with vision cones or topological interactions are studied in [4, 32, 67] and with noise [50, 61].

In the theorem below, we summarize the flocking estimates for the system (3). The proof follows the blueprint of [1, 40, 65, 85, 96], so we omit it here.

Theorem 1 *Let (x, v) be any global smooth solution to the CS system (3)–(4).*

- *If $0 \leq \beta \leq 1$, then we have unconditional asymptotic flocking, that is*

$$R^v(0)e^{-t} \leq R^v(t) \leq R^v(0)e^{-\psi(\tilde{R})t} \quad t \geq 0, \quad (6)$$

where \tilde{R} is implicitly given by

$$R^v(0) = \int_{R^x(0)}^{\tilde{R}} \psi(s) ds.$$

- *If $\beta > 1$ and the initial diameters $R^x(0)$ and $R^v(0)$ satisfy*

$$R^v(0) < \int_{R^x(0)}^{\infty} \psi(s) ds, \quad (7)$$

then estimate (6) also holds.

In Figs. 1 and 2, we observe typical particle simulations of the CS model in 2D. As stated in Theorem 1, the unconditional asymptotic flocking occurs for any initial data in the case simulated in Fig. 1 while Fig. 2 shows a comparison between flocking and non-flocking cases. In Fig. 1, the initial positions and velocities $\{(x_{i0}, v_{i0})\}_{i=1}^N$ with $N = 50$ are generated randomly from the uniform distribu-

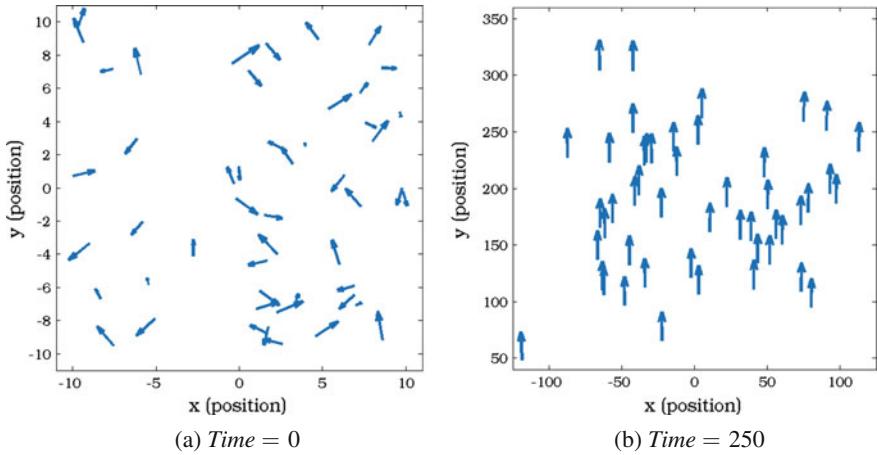


Fig. 1 Large-time behavior of solutions to the CS model with $\beta = 0.8$.

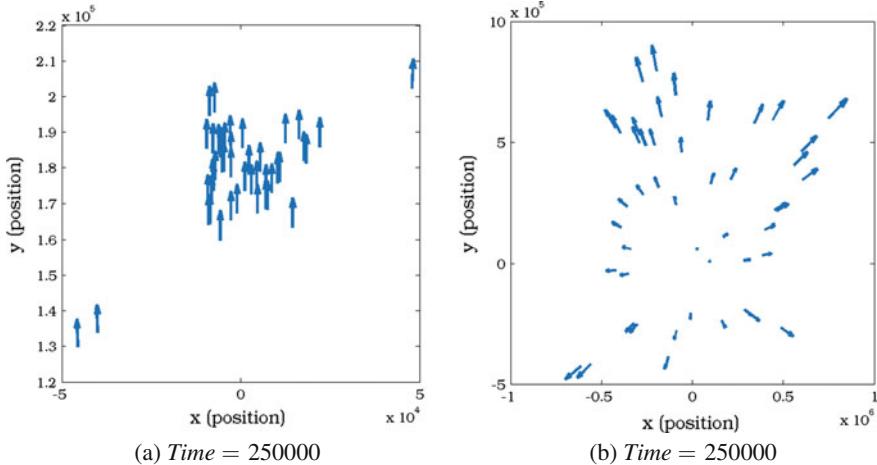


Fig. 2 Large-time behavior of solutions to the CS model with $\beta = 1.05$ (A) and $\beta = 1.2$ (B).

tion $[-10, 10]^2 \times \{[-5, 5] \times [-4.3, 5.7]\}$ with the aim of having the mean velocity equal to $\frac{1}{N} \sum_{i=1}^N v_i(0) \approx (0, 0.7)$.

In Fig. 2, the flocking behavior happens depending on the initial data. The initial positions and velocities $\{(x_{i0}, v_{i0})\}_{i=1}^N$ with $N = 50$ are generated randomly from the uniform distribution $[-10, 10]^2 \times \{[-5, 5] \times [-4.3, 5.7]\}$ with the aim of having the mean velocity equal to $\frac{1}{N} \sum_{i=1}^N v_i(0) \approx (0, 0.7)$. With this initial configuration, it results that $R^x(0) = 26.23$ and $R^v(0) = 12.25$. Then, the initial data for (A) satisfy (7) since $R^v(0) < \int_{R^x(0)}^\infty \psi(s) ds = 16.43$. On the other hand, the initial data for (B) do not satisfy the condition (7) because $R^v(0) > \int_{R^v(0)}^\infty \psi(s) ds = 2.60$.

Remark 1 For the CS model, i.e., $S_i(x) \equiv N$ in (3), if we set an averaged quantity $v_c(t) := \frac{1}{N} \sum_{i=1}^N v_i(t)$, then $v_c(t)$ satisfies $v'_c(t) = 0$, i.e., $v_c(t) = v_c(0)$ due to the symmetry of the communication function ψ . Thus, if the global flocking occurs, then we have that for all $i \in \{1, \dots, N\}$

$$v_i(t) \rightarrow v_c(0) = \frac{1}{N} \sum_{i=1}^N v_i(0) \quad \text{as } t \rightarrow \infty.$$

On the other hand, in the MT model, i.e., $S_i(x) = \sum_{k=1}^N \psi(x_i - x_k)$, the momentum is not conserved. Thus, identifying the asymptotic flocking state in terms of the initial data is a very intriguing question. Partial answers to asymptotic flocking have been provided in [85].

In Fig. 3, we show the different behavior of the CS and MT velocity averaging. We choose the initial positions $\{x_{i0}\}_{i=1}^{55}$ divided into two groups, $G_1 := \{x_{i0}\}_{i=1}^{50}$ and $G_2 := \{x_{i0}\}_{i=1}^5$, and they are generated randomly from the uniform distribution $[-10, 10]^2$ and $[60, 63] \times [-1.5, 1.5]$, respectively. The initial velocities $\{v_{i0}\}_{i=1}^{55}$ are generated randomly from the uniform distribution $[-5, 5] \times [-4.3, 5.7]$ with the aim of having the mean velocity equal to $\frac{1}{N} \sum_{i=1}^N v_i(0) \approx (0, 0.7)$. We observe the much faster decay of the velocity radius of the support $R^v(t)$ defined in (5) in the MT model compared to the CS model, and thus, the asymptotic flocking is achieved faster in the MT model than in the CS model.

Finally, we show in Fig. 4 a comparison of the time evolution of the system with $S_i(x) \equiv N$ and $S_i(x) = \sum_{k=1}^N \psi(x_i - x_k)$. Subplots (a) and (b) show a snapshot of the solutions at $t = 5$, while subplots (c) and (d) show a snapshot of the solutions at $t = 50$, for the CS and the MT models, respectively. From Fig. 4 (a), we find that

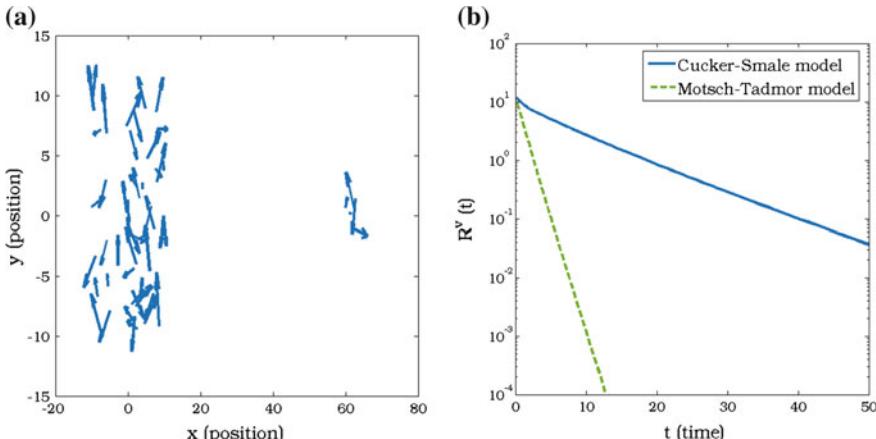


Fig. 3 (a): Initial positions are divided into two groups. (b): Comparison of the log scale decay rate of $R^v(t)$ for both systems.

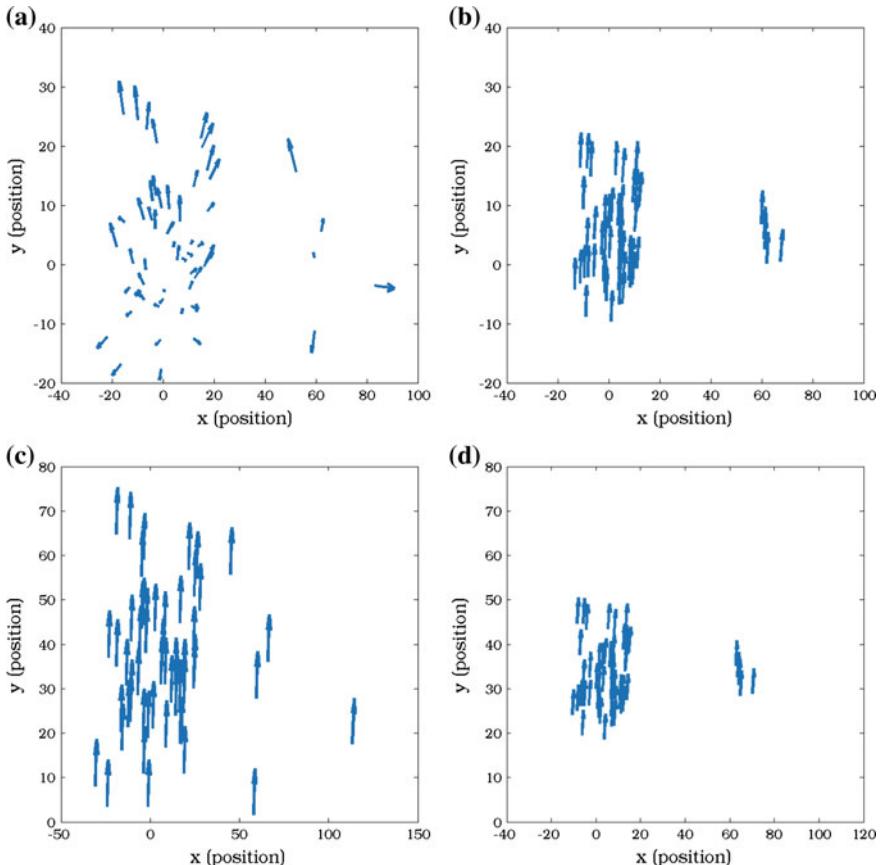


Fig. 4 Comparison of time evolution of the system with $S_i(x) \equiv N$ and $S_i(x) = \sum_{k=1}^N \psi(x_i - x_k)$. Subplots (a) and (b) show a snapshot of the solutions at $t = 5$, and subplots (c) and (d) show a snapshot of the solutions at $t = 50$, for the CS and the MT models, respectively.

the CS flocking particles in the small group G_2 are not interacting with others in the large group G_1 in the beginning. It seems that the particles tend to move with their own initial velocities. On the other hand, the particles in the MT model are trying to be aligned with their neighbors from the beginning. Even though both models exhibit flocking behavior and the analytical results require the same conditions for flocking, numerical simulations demonstrate again that the decay rate of convergence of the MT model to the flocking state is faster than the one of the CS models. We are not aware of results comparing the rate of decay to flocking for both models.

2.2 Attractive–Repulsive Models

Now, we turn to the case in which we incorporate attractive–repulsive forces to the system (1) with the CS alignment force:

$$\begin{cases} \dot{x}_i = v_i, & i = 1, \dots, N, \quad t > 0, \\ \dot{v}_i = \frac{1}{N} \sum_{j=1}^N \psi(x_i - x_j)(v_j - v_i) - \frac{1}{N} \sum_{j=1}^N \nabla K(x_i - x_j). \end{cases} \quad (8)$$

For the interaction potential K , we will choose in most of our simulations a repulsive Newtonian potential confined by a quadratic one of the forms

$$K(x) = \alpha \frac{|x|^2}{2} + k\phi(x) \quad \text{where} \quad -\Delta_x \phi(x) = \delta_0(x). \quad (9)$$

By choosing $\alpha > 0$, we confine our particles in a bounded region and they are repelled by Newtonian interaction choosing $k < 0$. In general, K is typically chosen repulsive at the origin and attractive at infinity in such a way that there is a typical length of stable interaction between two particles. Other popular choices as mentioned above are Morse-like and power law-like potentials as in [42, 43, 47, 60] and the references therein.

The existence of particular solutions, called flock solutions, for the system (8) has recently received lots of attention due to the ubiquitous appearance of this kind of solutions in several swarming models.

Definition 2 A flock solution of the particle model (8) is a spatial configuration \hat{x} with zero net interaction force on every particle, that is:

$$\sum_{j \neq i} \nabla K(\hat{x}_i - \hat{x}_j) = 0, \quad i = 1, \dots, N,$$

that translates at a uniform velocity $m_0 \in \mathbb{R}^d$, hence $(x_i(t), v_i(t)) = (\hat{x}_i - tm_0, m_0)$.

The richness of the qualitative properties of the spatial configurations for the flock solution, also called flock profile, depending on the potential K is quite impressive, see [76]. Other stable patterns were observed for related systems, for instance single or double rotating mills [38, 45, 46, 60, 79]. However, these milling patterns are typically eliminated due to the presence of the CS alignment term. The stability of flock patterns for the particle system (8) has recently been established in [3, 44].

As the total number of individuals gets large, the system of differential equations is difficult to analyze and usually a continuum description based on mean-field limits is adopted, either at the kinetic level for the particle distribution function [38, 41] or at the hydrodynamic level for the macroscopic density and velocities [38, 52] as we will discuss in the next section. At the continuum level, the flock profiles are characterized

by searching for continuous probability densities or probability measures ρ of particle locations such that the total force acting on each individual balances out. This is equivalent to finding probability densities or measures ρ such that

$$\nabla K * \rho = 0 \quad \text{on } \text{supp}(\rho). \quad (10)$$

Being the problem posed on the support of the unknown density ρ implies that the equation (10) is highly nonlinear. In fact, characterizing the interaction potentials K such that these profiles are continuous or regular in their support is a very challenging question. Explicit formulas for solutions to (10) for particular potentials such as Morse-like and power law-like potentials are possible due to the particular properties of associated differential operators [16, 42, 43, 47, 62, 80]. In particular, it is known from classical potential theory that the solution to (10) in the case of the confined repulsive Newtonian potential (9) is given by a characteristic of a ball whose radius is determined to have the right total mass of the system.

In Fig. 5, we show a typical simulation for the system (8) with interaction potential given in (9). The initial positions and velocities $\{(x_{i0}, v_{i0})\}_{i=1}^N$ with $N = 50$ are generated randomly from the uniform distribution $[-10, 10]^2 \times \{[-5, 5] \times [-4.3, 5.7]\}$ with the aim of having the mean velocity equal to $\frac{1}{N} \sum_{i=1}^N v_i(0) \approx (0, 0.7)$. This simulation shows the generic flock formation after some time in which particles distribute more or less uniformly in a certain ball. Of course, we know that as $N \rightarrow \infty$ this distribution of particles will be getting closer and closer to the characteristic of the ball, see [18, 21, 76] for instance.

Let us finally mention that the continuum spatial profile of the flock solutions can also be found by steepest descent methods from the first-order models of swarming.

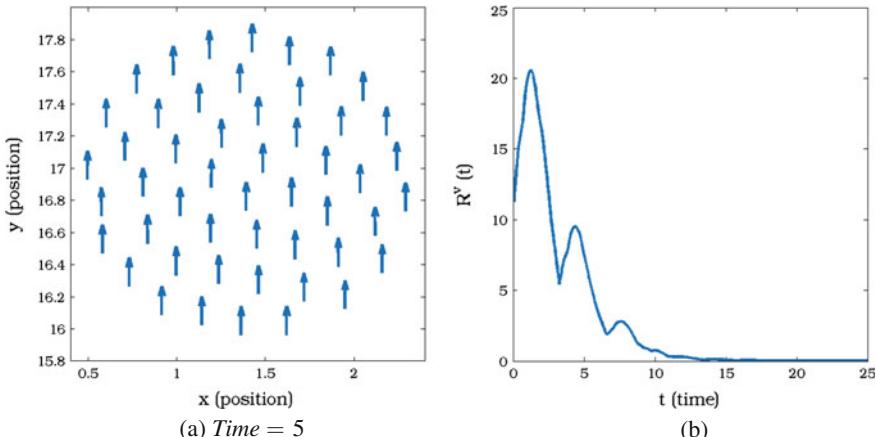


Fig. 5 Large-time behavior of solutions to the system (8) with $k = -1$, $\alpha = 1$, and $\beta = 0.5$. (a): The particle is uniformly distributed in the unit circle. (b): $R^v(t)$ eventually converges to zero as time goes on. However, it is not monotonically decreasing.

These models can be found formally from the previous second-order particle models by substituting the CS alignment term by simple linear friction force and assuming that inertia is negligible with respect to other terms, see [83, 84]. In this limit, they lead to

$$\frac{d}{dt}x_i = -\frac{1}{N} \sum_{j \neq i} \nabla K(|x_i - x_j|), \quad i = 1, \dots, N. \quad (11)$$

By taking the mean-field limit in (11), as $N \rightarrow \infty$, one derives the so-called aggregation equation

$$\rho_t = \nabla \cdot (\rho \nabla K * \rho), \quad (12)$$

for the evolution of the mass density of particles. The aggregation equation (12) with repulsive–attractive potentials has attracted lots of attention in the last years in the mathematical analysis community due to its rich regularity structure for steady states and solutions depending on the singularity of the potential at the origin, see [10–12, 17, 19–22, 30, 36, 37, 48, 49, 78, 97, 98] and the references therein.

3 Macroscopic Descriptions: Flocking Behavior and Finite-Time Blowup Phenomena

In this section, we study some of the features and properties of continuum models for collective behavior being capable of describing flocking behavior. By using BBGKY hierarchies or mean-field limits [41], we can derive a Vlasov-type equation from the system (8). More precisely, when the number of individuals goes to infinity, i.e., $N \rightarrow \infty$, the mesoscopic observables for the system (8) can be calculated from the velocity moments of the density function $f = f(x, v, t)$ which is a solution to the following Vlasov-type equation:

$$\begin{cases} \partial_t f + v \cdot \nabla_x f + \nabla_v \cdot (F(f)f - (\nabla_x K \star \rho)f) = 0, & x, v \in \mathbb{R}^d, t > 0, \\ F(f) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \psi(x - y)(w - v)f(y, w, t) dy dw, \\ \rho(x, t) = \int_{\mathbb{R}^d} f(x, v, t) dv. \end{cases} \quad (13)$$

The derivation of the kinetic equation (13) is well studied only for regular potentials [25, 59, 87]. If K vanishes, rigorous mean-field limit, existence of weak solutions, and large-time behavior of measure-valued solutions are studied in [30, 40, 65]. We also refer to [6–9, 33, 51] for a dynamics of flocking particles interacting with homogeneous/inhomogeneous fluids. For the equation (13) under certain conditions for K , quite general frameworks are proposed in [23, 29, 31]. For not too singular interaction potentials K , the rigorous derivation of (13) is studied in [69].

The kinetic description has to be taken as usual as an intermediate mesoscopic description of the system leading to macroscopic models for collective behavior via asymptotic limits or closure assumptions. By taking moments on the kinetic equation (13) together with a zero temperature closure or monokinetic assumption for the local hydrodynamics solution, one can obtain hydrodynamic descriptions of the system (8). This procedure has been performed in different ways by different authors, see for instance [38, 52, 66]. Associated with the kinetic distribution function $f(x, v, t)$, one can define the mean velocity as

$$\rho(x, t)u(x, t) = \int_{\mathbb{R}^d} vf(x, v, t) dv.$$

By taking the first two moments with respect to v on the kinetic equation (13), one formally obtains

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0, & x \in \mathbb{R}^d, \quad t > 0, \\ \partial_t(\rho u) + \nabla_x \cdot (\rho u \otimes u) + \nabla_x \cdot \left(\int_{\mathbb{R}^d} (v - u) \otimes (v - u) f(x, v, t) dv \right) = \\ \quad \rho \int_{\mathbb{R}^d} \psi(x - y)(u(y) - u(x))\rho(y) dy - \rho(\nabla_x K \star \rho). \end{cases} \quad (14)$$

Assuming that the distribution function is not far from monokinetic, that is

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

the hydrodynamic system (14) is reduced to the following pressureless Euler-type equations given by

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0, & x \in \mathbb{R}^d, \quad t > 0, \\ \partial_t(\rho u) + \nabla_x \cdot (\rho u \otimes u) = \rho \left(\int_{\mathbb{R}^d} \psi(x - y)(u(y) - u(x))\rho(y) dy - \nabla_x K \star \rho \right), \end{cases} \quad (15)$$

where $\rho = \rho(x, t)$ and $u = u(x, t)$ represent the particle density and their corresponding mean velocity, respectively. For now on, $\Omega(t)$ denotes the interior of the support of the density ρ , i.e., $\Omega(t) := \{x \in \mathbb{R} : \rho(x, t) > 0\}$. We assume that $\Omega(0) =: \Omega_0$ is a bounded open set. The hydrodynamic system (15) has to be complemented with initial conditions

$$(\rho(\cdot, t), u(\cdot, t))|_{t=0} = (\rho_0, u_0). \quad (16)$$

As usual with hydrodynamic equations, we observe that they are mathematically challenging due to the nonlinearity introduced by the material derivative of the velocity field that can lead to blowup of the velocity profile. On the other hand, flock profiles are also solutions of the hydrodynamic equations (15). Actually, if the

density ρ satisfies (10) and the velocity is constantly given by $u(x, t) = u_0 \in \mathbb{R}^d$, then they form a particular solution of the hydrodynamic equations (15). This is the flocking solution at the hydrodynamical level of description for collective behavior.

In the next subsections, we will numerically explore the derived hydrodynamic equations (15) in one dimension providing numerical evidence showing the flocking behavior and the finite-time blowup of solutions giving some insights into the stability of flock solutions and the conditions for blowup of the solutions. Since analytical results concerning these hydrodynamic equations are very few in the literature, we will compare to existing analytical results in the relevant sections below.

3.1 Numerical Scheme

For the numerical simulation of the hydrodynamic system (15), we use a Lagrangian numerical scheme. With this purpose, we consider the characteristic flow $\eta(x, t)$ associated with the fluid velocity u defined by

$$\frac{d\eta(x, t)}{dt} = u(\eta(x, t), t) =: v(x, t) \quad \text{with} \quad \eta(x, 0) = x. \quad (17)$$

Set $h(x, t) := \rho(\eta(x, t), t)$, then using the characteristic flow (17), we can rewrite the system (15) in one dimension as

$$\begin{cases} h(x, t) = \rho_0(x) \left(\frac{\partial \eta}{\partial x}(x, t) \right)^{-1}, \\ \frac{dv}{dt}(x, t) = \int_{\mathbb{R}} \psi(\eta(x, t) - \eta(y, t)) (v(y, t) - v(x, t)) \rho_0(y) dy \\ \quad - \int_{\mathbb{R}} \frac{\partial K}{\partial x}(\eta(x, t) - \eta(y, t)) \rho_0(y) dy, \end{cases} \quad (18)$$

with the initial data

$$(h(\cdot, t), v(\cdot, t))|_{t=0} = (\rho_0, u_0).$$

Note that the continuity equation for the density (18)₁ is decoupled from the rest. Thus, it can easily be solved from the information obtained from the characteristic and momentum equations.

We do a spatial discretization of the equation (18)₂ choosing a uniform mesh of length Δx of the initial positions of the particles with a number of points given by n . For each node i , the position, density, and velocity through the characteristics of the i th-particle will be denoted as $\eta_i(t)$, $h_i(t)$ and $v_i(t)$. At each node, the term $\frac{\partial \eta_i(t)}{\partial x}$ is computed from $\eta_j(t)$, that is with the information about how the position of the nodes changes through the characteristics in time, by standard finite differences of fourth order, and afterward that value is inverted. The end values use one-sided

finite differences to avoid unknown values of the Lagrangian density. The spatial derivative of the interaction potential appearing in (18)₂ is obtained analytically and not approximated numerically. For the integral terms in (18)₂, we approximate them by direct numerical quadrature formulas as

$$\begin{aligned} \int_{\Omega_0} \psi(\eta(x, t) - \eta(y, t)) v(y, t) \rho_0(y) dy &\sim \Delta x \sum_{j=1, j \neq i}^n \psi(\eta_i(t) - \eta_j(t)) v_j(t) \rho_j(0), \\ \int_{\Omega_0} \psi(\eta(x, t) - \eta(y, t)) \rho_0(y) dy &\sim \Delta x \sum_{j=1, j \neq i}^n \psi(\eta_i(t) - \eta_j(t)) \rho_j(0), \\ \int_{\Omega_0} \frac{\partial K}{\partial x}(\eta(x, t) - \eta(y, t)) \rho_0(y) dy &\sim \Delta x \sum_{j=1, j \neq i}^n \frac{\partial K}{\partial x}(\eta_i(t) - \eta_j(t)) \rho_j(0). \end{aligned}$$

Taking everything into consideration, the system of equations (18), for each node i , takes the form

$$\begin{cases} \frac{d\eta_i(t)}{dt} = v_i(t), \\ h_i(t) = \rho_i(0) \left(\frac{\partial \eta_i(t)}{\partial x} \right)^{-1}, \\ \frac{dv_i(t)}{dt} = \Delta x \sum_{j=1, j \neq i}^n \rho_j(0) \left[\psi(\eta_i(t) - \eta_j(t)) (v_j(t) - v_i(t)) - \frac{\partial K}{\partial x}(\eta_i(t) - \eta_j(t)) \right]. \end{cases} \quad (19)$$

A temporal discretization is carried out in the system composed by the Equations (19)₁ and (19)₃, in order to obtain the evolution of position and velocity through the characteristics. The employed numerical scheme is a classical fourth-order Runge–Kutta explicit scheme using the built-in *ode45* MATLAB command. Subsequently, the density is obtained from (19)₂. The numerical experiments have been performed with the version R2014a of MATLAB, using a computer ACER Aspire V3-572G. Similar Lagrangian approaches have been taken by other authors in related problems [45, 74].

In all the test cases below, we choose the same initial conditions for the particle positions, and the initial position of the nodes has been set uniformly distributed inside the interval $[-0.75, 0.75]$. Thus, the initial position of each node i is given by

$$\eta_i(0) = -0.75 + \frac{1.5}{n-1} (i-1) \quad \text{for } i = 1, \dots, n.$$

In most of our simulations $n = 200$ if it is not specified otherwise. The initial conditions of density and velocity will be specified for each case that will be treated subsequently.

3.2 Euler-Alignment System

The aim of this subsection is to analyze numerically some of the open problems related to the theoretical results studied in [34] on critical threshold phenomena for the system (15) without the interaction forces. We will take the communication function of CS model given by

$$\psi(x) = \frac{1}{(1 + |x|^2)^{\beta/2}}, \quad \beta > 0.$$

In this case, the global regularity or the finite-time blowup of the solution can be determined by the initial configurations with sharp conditions. We use this case as validation to our scheme being capable of showing either the global consensus in velocity or the blowup in the velocity field and density as shown by the theory [34]. We also numerically compare the large-time behaviors of solutions to the CS and the MT models at the hydrodynamic level. The simulations in this subsection are done with initial density

$$\rho_i(0) = \frac{1}{\gamma} \cos \left(\pi \frac{x_i(0)}{1.5} \right) \quad \text{for each node } i = 1, \dots, n,$$

where the constant γ is fixed by the mass normalization, i.e., $\int_{\mathbb{R}} \rho_0 dx = 1$. Concerning the initial velocity, we choose

$$u_i(0) = -c \sin \left(\pi \frac{x_i(0)}{1.5} \right) \quad \text{for each node } i = 1, \dots, n,$$

where the constant $c > 0$ will be varied to study different initial conditions in the simulations except for the comparison between the MT and CS models.

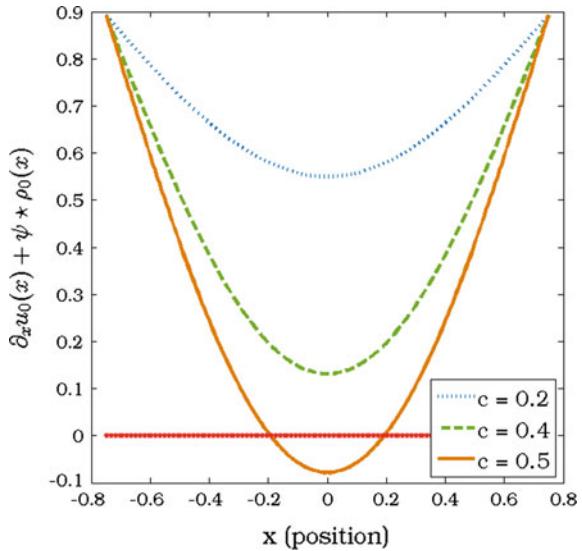
3.2.1 Hydrodynamic Cucker–Smale Model

In this case, a critical threshold leading to sharp dichotomy between global-in-time existence and finite-time blowup of solutions is provided in [34]. The region where the solutions blow up in a finite time is called “*Supercritical region*”; otherwise, the region in which the solutions globally exist in time is called “*Subcritical region*.”

Theorem 2 *Let (ρ, u) be classical solutions to the system (15)–(16) with $K = 0$.*

- (*Subcritical region*) *If $\partial_x u_0(x) \geq -\psi \star \rho_0(x)$ for all $x \in \mathbb{R}$, then the system has a global classical solution.*
- (*Supercritical region*) *If there exists an x such that $\partial_x u_0(x) < -\psi \star \rho_0(x)$, then there is a finite-time blowup of the solution. Moreover, this blowup happens as an infinite negative slope in the velocity and divergence value of the density at the same location.*

Fig. 6 Values of $\partial_x u_0(x) + \psi * \rho_0(x)$ for different values of c .



For the numerical simulations, three different cases corresponding with the values of the constant $c = 0.2, 0.4, 0.5$ are treated. The first two cases lie in subcritical region, and the third one lies in the supercritical region, see Fig. 6.

Since the initial configurations of the first two cases lie in the subcritical region, we have global regularity of solutions. Both initial configurations are symmetric; thus, the initial mean velocity and center of mass are zero, which are kept through their evolution. The numerical simulations in Fig. 7 demonstrate that they are consistent with Theorem 2 leading to global-in-time solutions. Even though both cases, $c = 0.2, 0.4$, are inside the subcritical region, the steady density for the case $c = 0.4$ shows more concentrated behavior at the middle of the domain (Fig. 7. (c)) than the case $c = 0.2$ (Fig. 7. (a)). For both cases, the velocity converges to zero as time goes on which gives the global consensus or flocking behavior in this case. We see that the profile of density depends in a complicated way on the initial density configuration as it happens in the particle system.

When $c = 0.5$, the initial data lie in the supercritical region, and thus, a blowup should be expected from Theorem 2. Indeed, the numerical scheme is capable to show this blowup phenomenon. We can observe that the derivative of the velocity becomes sharper and sharper at the origin, see the inlet in Fig. 8(b), while the density value at the origin gets larger and larger, Fig. 8(a). Actually, the simulation cannot be continued after $t = 2.7311$ due to the high value of the density and the large negative derivated of the velocity at the origin. Fig. 8 depicts the time–behavior of density until $t = 2$.

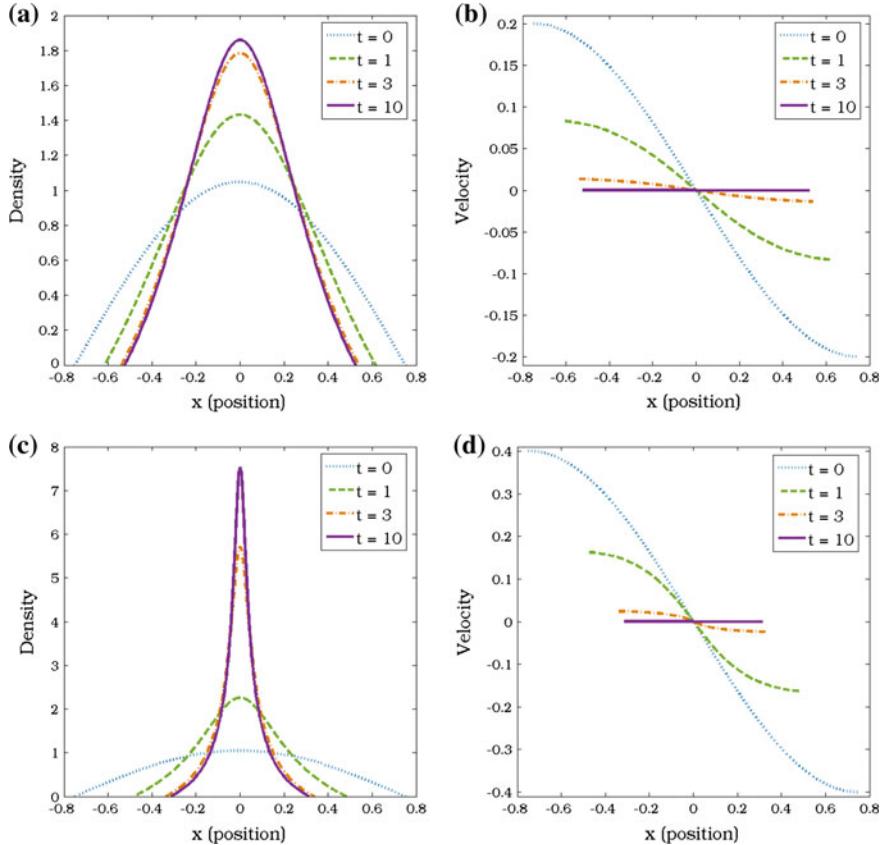


Fig. 7 Subcritical Region: Global Existence and Asymptotic Flocking.- (a), (b): The evolution of density and velocity for the case of $c = 0.2$. (c), (d): The evolution of density and velocity for the case of $c = 0.4$.

3.2.2 Numerical Comparison of Cucker–Smale and Motsch–Tadmor Equations

The aim of this subsection is to compare the solutions of the hydrodynamic CS and MT systems. Taking a similar strategy for the CS model, the hydrodynamic MT system can be formally derived from the kinetic and the particle levels of the MT model, leading to

$$\begin{cases} \partial_t \rho + \partial_x (\rho u) = 0, & x \in \mathbb{R}, \quad t > 0, \\ \partial_t (\rho u) + \partial_x (\rho u^2) = \frac{1}{\psi * \rho} \int_{\mathbb{R}} \psi(x-y)(u(y)-u(x))\rho(x)\rho(y) dy. \end{cases} \quad (20)$$

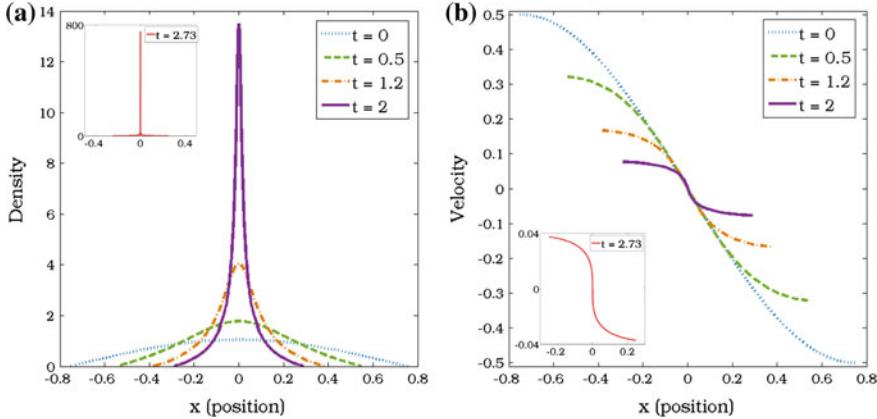


Fig. 8 Supercritical Region. The evolution of density (a) and velocity (b) for the case of $c = 0.5$.

For the system (20), the large-time behavior of solutions showing the velocity alignment is studied in [85] and the critical threshold phenomenon is provided in [95].

For the numerical simulations, we take the parameter $\beta = 0.5$ in the communication function ψ . The initial conditions have been chosen following the simulations carried out at the particle level, see Figs. 3 and 4. The objective has been to establish two zones in the initial domain with an important difference in mass. In addition, those two regions start with opposite velocity, so that the direction of the velocity corresponding to the largest zone will prevail.

The initial distances of Fig. 3 have been preserved in the simulation, although here they have been divided by 10 for visualization. Then, the initial positions of the nodes are

$$\eta_i(0) = -1 + \frac{7.5}{n-1} (i-1) \quad \text{for } i = 1, \dots, n.$$

Concerning the initial density, the sum of the mass of both regions is unit, and the relation between the individual masses is the same as in the particle level, 50/5. Then the initial density is a piecewise function that satisfies

$$\rho_i(0) = \begin{cases} \frac{1}{\gamma_1} \cos \left(\pi \frac{x_i(0)}{2} \right) & \text{if } \eta_i(0) \in [-1, 1], \\ 0 & \text{if } \eta_i(0) \in (1, 5.5), \\ \frac{1}{\gamma_2} \cos \left(\pi \frac{x_i(0) - 6}{1} \right) & \text{if } \eta_i(0) \in [5.5, 6.5] \end{cases} \quad \text{for } i = 1, \dots, n,$$

where γ_1 and γ_2 are chosen to satisfy the conditions on the masses. With respect to the initial velocities, both groups start with opposite velocities given by

$$u_i(0) = \begin{cases} 0.1 \cos\left(\pi \frac{x_i(0)}{2}\right) & \text{if } \eta_i(0) \in [-1, 1], \\ 0 & \text{if } \eta_i(0) \in (1, 5.5), \\ -0.1 \cos\left(\pi \frac{x_i(0) - 6}{1}\right) & \text{if } \eta_i(0) \in [5.5, 6.5] \end{cases} \quad \text{for } i = 1, \dots, n.$$

In Fig. 9 (a) and (b), the evolution of the density and velocity is showed, at $t = 20$. As it happened in the microscopic case, for the CS system the small group tends to keep the initial configuration, while for the MT case it varies more.

In order to compare the convergence rate to steady states of these systems, we consider the following quantity which measures the difference between velocities on the support of density:

$$R_\rho^v(t) := \sup_{x, y \in \Omega(t)} |u(x, t) - u(y, t)|.$$

As depicted in Fig. 9. (c), the MT model shows faster decay rate than the CS model, which is already observed at the particle level, see Fig. 3. (b).

3.3 Attractive–Repulsive Models

In this subsection, we consider the system (15) with the following power law potential in one dimension:

$$K(x) = \frac{|x|^a}{a} - \frac{|x|^b}{b},$$

with the convention that $|x|^0/0 = \log|x|$. In one dimension, $K(x) = k|x|$ corresponds to the attractive ($k > 0$) or repulsive ($k < 0$) Poisson force. In Section 3.3.1, we deal with this case and study the critical thresholds analytically and numerically. In particular, there is a gap between the known subcritical and supercritical regions for the repulsive Poisson force (see Theorem 3); we numerically analyze initial data in this region. In Section 3.3.2, we study the case $a = 2$ and $b = 1$. In this case, it is well known, as discussed in Section 2.2 and at the beginning of Section 3, that there is a flocking solution given by a density profile satisfying (10) whose solution is the characteristic function of an interval. In particular, there is a steady state with $u = 0$ with this density profile. We numerically show the convergence in time of density to this steady state whenever solutions are globally defined. We also study the blowup phenomena in this case which is not presented in the first-order models of swarming such as the aggregation equation (12). Finally, the cases $a = 2$ and $b = 0$, i.e., when the repulsive force is more repulsive than Newtonian, are also numerically explored in Section 3.3.3. In this case, the flocking profile/stationary state for the density satisfying (10) is given by the semicircle law as shown in [39] and the references therein.

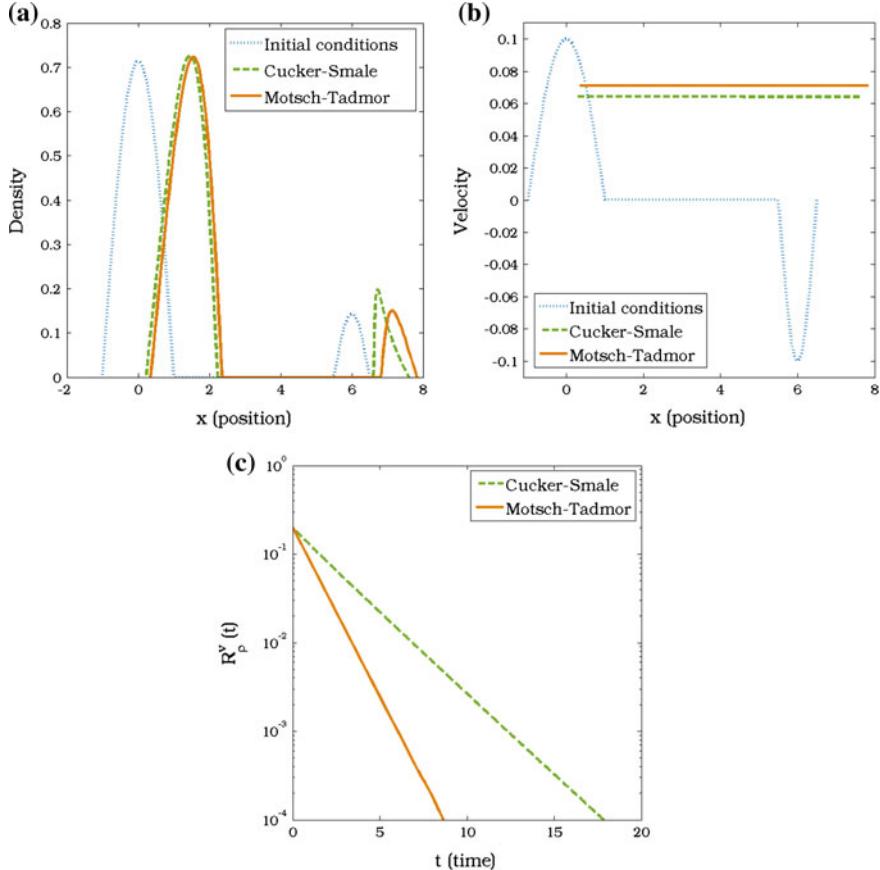


Fig. 9 Comparison between CS (15) with $K = 0$ and MT (20) systems. (a): Evolution of the density at $t = 20$. (b): Evolution of the velocity at $t = 20$. (c): Decay rate of the velocity.

3.3.1 Euler–Poisson–Alignment System

In this part, we consider the system (15) with Poisson forcing term in one dimension:

$$\begin{cases} \partial_t \rho + \partial_x (\rho u) = 0, & x \in \mathbb{R}, \quad t > 0, \\ \partial_t (\rho u) + \partial_x (\rho u^2) = \int_{\mathbb{R}} \psi(x-y)(u(y)-u(x))\rho(x)\rho(y) dy - \rho (\partial_x K \star \rho), \end{cases} \quad (21)$$

with $K(x) = k|x|$ attractive when $k > 0$ and repulsive when $k < 0$. Note that $k = \pm 0.5$ corresponds to the Newtonian interaction. For the system (21), the critical thresholds are studied in [34] depending on the sign of k . The result in [34] is as follows.

Theorem 3 Let (ρ, u) be solutions to the Euler-Poisson-alignment model (21).

1. Attractive potential ($k > 0$): A unconditional finite-time blowup of the solution will appear no matter the initial conditions.
 2. Repulsive potential ($k < 0$): As in the Euler-alignment system, there are two different zones:
- (Subcritical region) If $\partial_x u_0(x) < -\psi \star \rho_0(x) + \sigma_+(x)$ for all $x \in \mathbb{R}$, then the system has a global classical solution. Here, $\sigma_+(x) = 0$ whenever $\rho_0(x) = 0$ and elsewhere $\sigma_+(x)$ is the unique negative root of the equation

$$\rho_0^{-1}(x) - \frac{1}{\psi_M^2} \left(2k + \frac{\psi_M \sigma_+(x)}{\rho_0(x)} - 2ke^{\frac{\psi_M \sigma_+(x)}{2k\rho_0(x)}} \right) = 0, \quad \rho_0(x) > 0. \quad (22)$$

- (Supercritical region) If there exists an x such that

$$\partial_x u_0(x) < -\psi \star \rho_0(x) + \sigma_-(x), \quad \sigma_- := -\sqrt{-4k\rho_0(x)}$$

then the solution blows up in a finite time.

Note that there is a gap between the thresholds in the case of repulsive potential due to the nonlocality of the velocity alignment force. If we choose the constant communication function $\psi \equiv 1$, we can close this gap, see Corollary 1.

For the initial density and velocity for the numerical simulations, we take them as in Section 3.2: for each node $i = 1, \dots, n$,

$$\rho_i(0) = \frac{1}{\gamma} \cos \left(\pi \frac{x_i(0)}{1.5} \right) \quad \text{and} \quad u_i(0) = -c \sin \left(\pi \frac{x_i(0)}{1.5} \right).$$

Similarly as before, we change the values of c to consider the subcritical and supercritical regions stated in Theorem 3.

Fig. 10 shows the evolution of density and velocity for the system (21) with $c = 0.4$ and $k = 0.5$. As stated in Theorem 3, the density is blowing up at a finite time in this case. It occurs when $t = 1.0788$.

For the repulsive potential case, we fix the value $k = -0.5$ and change the parameter $c = 0.95, 1.08, 1.2$ to consider both subcritical and supercritical regions described in Theorem 3, see Fig. 11. To be more precise, when $c = 0.95$, $\partial_x u_0(x) + \psi \star \rho_0(x)$ lies in the subcritical region, for $c = 1.08$ it is between subcritical and supercritical regions. Thus, it is not clear to have global regularity or finite-time blowup of solutions for that case. When $c = 1.2$, it is inside the supercritical region and finite-time blowing up solution is expected from Theorem 3. Since the initial density is independent of the parameter c , σ_- and σ_+ are same for all three cases.

A trust region with the Dogleg method is used in (22) to obtain the value of σ_+ . Basically it consists in a Newton–Raphson method that applies a trust-region technique to improve robustness when starting far from the solution. Additionally, it

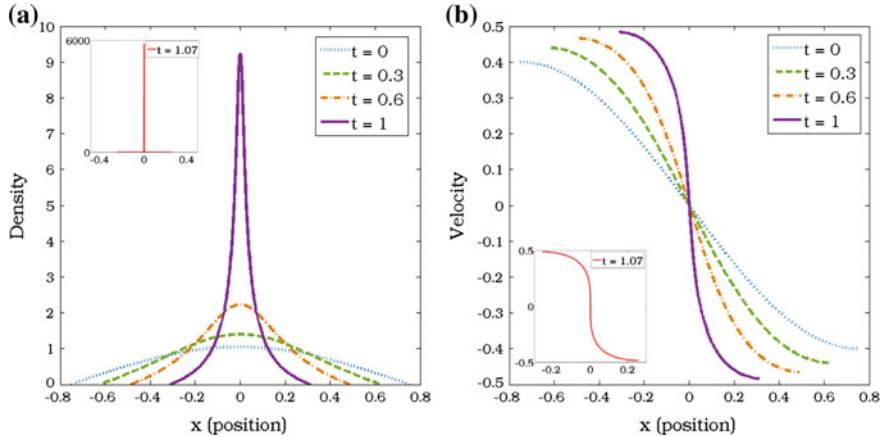
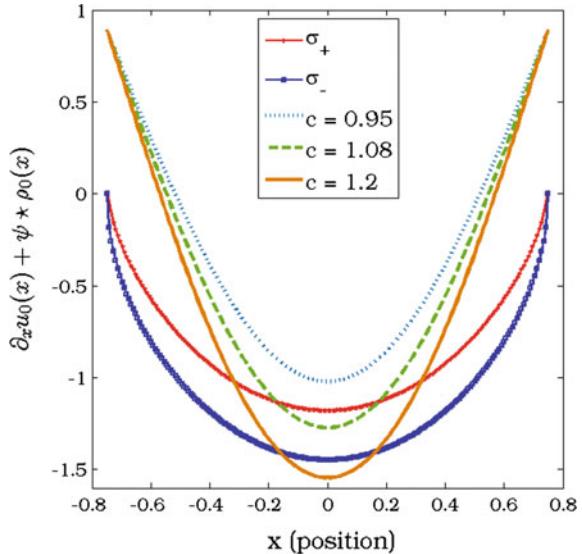


Fig. 10 Evolution of the density (a) and velocity (b) for the system (21) with $k = 0.5$ and $c = 0.4$.

Fig. 11 Value of $\partial_x u_0(x) + \psi * \rho_0(x)$ depending on c .



introduces a procedure denoted as Powell dogleg. For more information about this method, see [90]. In MATLAB, it could be easily implemented by the subroutine *fsolve*.

The numerical simulations of the density and the velocity for the case $c = 0.95$ at different times are shown in Fig. 12. In the beginning, the density tends to be concentrated near the origin, but after some time the repulsive force changes the sign of slope of the velocity. Consequently, the density spreads and the size of support of the density is increasing. Thus, there is no finite-time blowup of solutions in this case, which is consistent with Theorem 3.

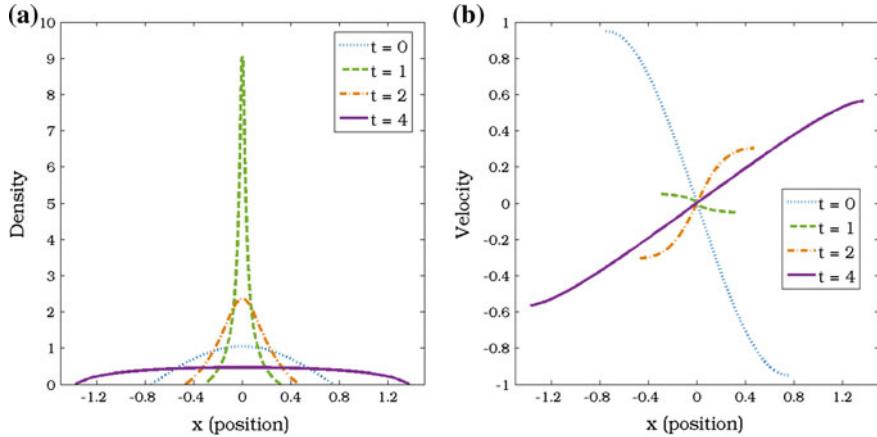


Fig. 12 Numerical simulation of the density (A) and velocity (B) for the case of $k = -0.5$ and $c = 0.95$.

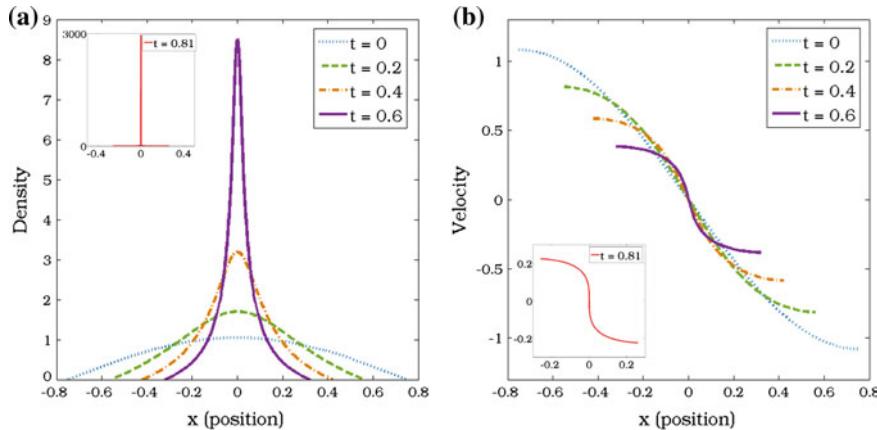


Fig. 13 Numerical simulation of the density (A) and velocity (B) for the case of $k = -0.5$ and $c = 1.08$.

When $c = 1.08$, some values of $\partial_x u_0(x) + \psi \star \rho_0(x)$ are between $\sigma_+(x)$ and $\sigma_-(x)$. We do not know analytically what to expect in this case. We numerically observe finite-time blowup in Fig. 13 after $t = 0.8107$. Our numerical exploration did not find initial data leading to global regularity when some values of $\partial_x u_0(x) + \psi \star \rho_0(x)$ are between $\sigma_+(x)$ and $\sigma_-(x)$. It is an open problem to decide whether dichotomy of coexistence of finite-time blowup and global existence can happen in this region.

Finally, when $c = 1.2$, the initial conditions are inside the supercritical region, and it is expected from Theorem 3 the blowup of solutions in finite time. The numerical simulations in this case show indeed the blowup behavior of solutions, see Fig. 14.

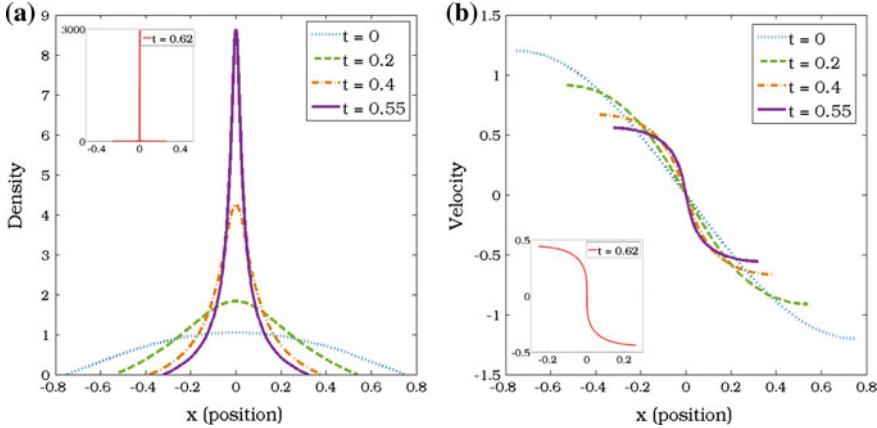


Fig. 14 Numerical simulation of the density (A) and velocity (B) for the cases $k = -0.5$ and $c = 1.2$.

The repulsive force is not enough to prevent the blowup phenomena of the system (21) with these initial data, and the blowup is produced after $t = 0.6204$.

As mentioned before, the gap between thresholds occurs due to the nonlocality of the velocity alignment force. In fact, if we choose the constant communication function, i.e., $\psi \equiv 1$ or $\beta = 0$, then we have a sharp critical threshold for the repulsive potential case.

Corollary 1 *Let (ρ, u) be solutions to the system (21) with $\psi \equiv 1$ and $k < 0$.*

- (*Subcritical region*) *If $\partial_x u_0(x) > -\|\rho_0\|_{L^1} + \sigma(x)$ for all $x \in \mathbb{R}$, then the system has a global classical solution. Here, $\sigma(x) = 0$ whenever $\rho_0(x) = 0$ and elsewhere $\sigma(x)$ is the unique negative root of the equation*

$$\rho_0^{-1}(x) - 2k - \frac{\sigma(x)}{\rho_0(x)} + 2ke^{\frac{\sigma(x)}{2k\rho_0(x)}} = 0, \quad \rho_0(x) > 0.$$

- (*Supercritical region*) *If there exists an x such that $\partial_x u_0(x) < -\|\rho_0\|_{L^1} + \sigma(x)$, where the value of $\sigma(x)$ is the one given in the subcritical region, then the solution blows up in a finite time.*

This particular case has also been checked to validate our code. In fact, our code captures the dichotomy in this case quite nicely leading to similar simulations as in the case of non-constant communication function.

3.3.2 Euler–Poisson–Alignment System with Power Law Potential

In this part, we consider the repulsive Newtonian potential confined by a quadratic attractive potential:

$$K(x) = -\frac{|x|}{2} + \frac{x^2}{2}. \quad (23)$$

For simplicity, we also consider a linear damping in the momentum equation instead of the nonlocal velocity alignment. In this situation, our main system reads as

$$\begin{cases} \partial_t \rho + \partial_x (\rho u) = 0, & x \in \mathbb{R}, \quad t > 0, \\ \partial_t (\rho u) + \partial_x (\rho u^2) = -\rho u - \rho (\partial_x K \star \rho). \end{cases} \quad (24)$$

If we take the constant communication function $\psi \equiv 1$ and assume the initial momentum is zero, i.e., $\int_{\mathbb{R}} \rho_0(x) u_0(x) dx = 0$, the system (24) can be derived from the system (15) with the potential K given in (23).

Concerning the initial density for the numerical simulations, we again define it, for each node $i = 1, \dots, n$, as

$$\rho_i(0) = \frac{1}{\gamma} \cos \left(\pi \frac{x_i(0)}{1.7} \right),$$

where the constant γ is fixed so that the total mass $M_0 := \int_{\mathbb{R}} \rho_0 dx$ has the required value. For the initial velocity, we choose

$$u_i(0) = -c x_i(0) \quad \text{for each node } i = 1, \dots, n,$$

where we again choose different constants $c > 0$ to deal with the subcritical and supercritical regions.

For the system (24) with the potential K given in (23), the sharp critical thresholds are classified in [35] according to the size of initial mass M_0 . Time-asymptotic behaviors of density and velocity are also studied. The results in [35] for the case $M_0 < 1/4$ are as follows.

Theorem 4 *Let (ρ, u) be a classical solution to the system (24) with the potential (23). Suppose that the initial density is compactly supported with the initial mass satisfying $M_0 < 1/4$. Then the solution blows up in finite time if and only if there exists a $x^* \in \Omega_0 := \text{supp}(\rho_0)$ such that*

$$\partial_x u_0(x^*) < 0, \quad M_0 - \rho_0(x^*) < \lambda_1 \partial_x u_0(x^*),$$

and

$$\rho_0(x^*) \leq (\lambda_1 \partial_x u_0(x^*) + \rho_0(x^*) - M_0)^{-\lambda_2/\sqrt{\Xi}} (\lambda_2 \partial_x u_0(x^*) + \rho_0(x^*) - M_0)^{\lambda_1 \sqrt{\Xi}}.$$

Here the constants $\lambda_i < 0$, $i = 1, 2$ and $\Xi > 0$ are given as

$$\lambda_1 := \frac{-1 + \sqrt{1 - 4M_0}}{2}, \quad \lambda_2 := \frac{-1 - \sqrt{1 - 4M_0}}{2}, \quad \text{and} \quad \Xi := 1 - 4M_0.$$

Furthermore, if there is no finite-time blowup, we have

$$\rho(x, t) \rightarrow M_0 \mathbf{1}_{[a, b]} \text{ in } L^1 \text{ and } u(x, t) \rightarrow 0, \text{ in } L^\infty \text{ as } t \rightarrow \infty,$$

exponentially fast, where a, b are constants given by

$$\begin{aligned} b &= \frac{1}{M_0} \left(\int_{\mathbb{R}} x \rho_0(x) dx + \int_{\mathbb{R}} (\rho_0 u_0)(x) dx \right) + \frac{1}{2}, \\ a &= \frac{1}{M_0} \left(\int_{\mathbb{R}} x \rho_0(x) dx + \int_{\mathbb{R}} (\rho_0 u_0)(x) dx \right) - \frac{1}{2}. \end{aligned}$$

We refer to [35] for the sharp critical thresholds and the large-time behavior for global-in-time solutions for $M_0 \geq 1/4$.

In order to check numerically the critical thresholds stated in Theorem 4, two numerical simulations are carried out in Fig. 15. First, the mass is set to be $M_0 = 0.2$, and then two cases $c = 0.9, 1.1$ are considered. Note that our initial conditions for ρ_0 and u_0 imply

$$\int_{\mathbb{R}} x \rho_0(x) dx = \int_{\mathbb{R}} (\rho_0 u_0)(x) dx = 0.$$

When $c = 0.9$, the initial data lie in the subcritical region; that is, the initial data do not satisfy the conditions in Theorem 4, and subsequently, the density and the velocity converge to $M_0 \mathbf{1}_{[-1/2, 1/2]}$ and 0 as time goes on, respectively. On the other hand, for the case $c = 1.1$, there is a finite-time blowup caused by an infinite slope of the velocity on the boundary. No numerical simulation has been provided for the case of $M_0 \geq 1/4$ because the critical threshold established in [35] for that case involves a more involved requirement on the initial conditions. However, the dichotomy of behaviors obtained is similar.

By employing the argument proposed in [34], we provide an estimate on the blowup time for the system (24) with the potential (23). It is worth mentioning that this blowup analysis does not depend on the size of mass. Differentiating the momentum equation of system (24) with respect to x , we can rewrite it as

$$\begin{cases} \rho' = -\rho d, \\ d' = -d^2 - d + 2\rho - M_0, \end{cases} \quad (25)$$

where $\{\cdot\}'$ denotes the time derivative along the characteristic flow η defined in (17) and $d := \partial_x u$.

Theorem 5 Let (ρ, u) be a classical solution to the system (24) with the potential (23) on the time interval $[0, T]$. Suppose that there exists an x such that

$$d_0(x) < 0 \text{ and } \frac{1 + d_0(x)}{\rho_0(x)} + 2 \log \left(1 - \frac{d_0(x)}{2\rho_0(x)} \right) \leq 0.$$

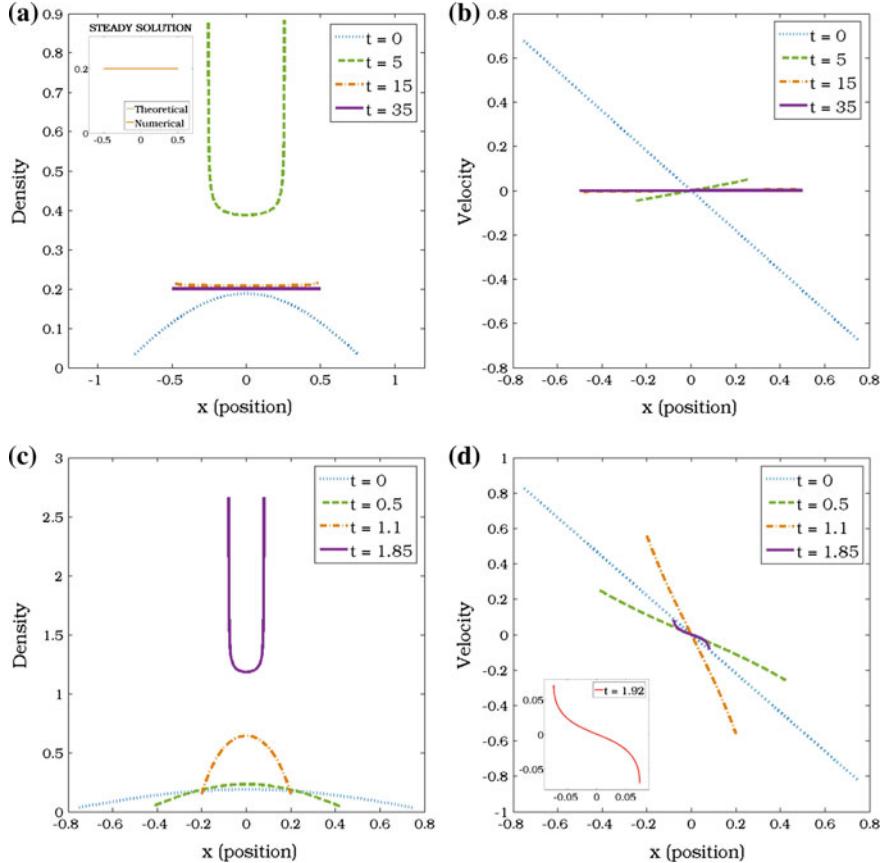


Fig. 15 Numerical simulations with respect to Theorem 4. (a), (b): Time behavior of the density and the velocity for the case $c = 0.9$. (c), (d): Time behavior of the density and the velocity for the case $c = 1.1$.

Then the life span T of the solution should be finite time and moreover

$$T \leq 2 \inf_{x \in \mathcal{S}} \log \left(1 - \frac{d_0(x)}{2\rho_0(x)} \right),$$

where \mathcal{S} is defined as

$$\mathcal{S} := \left\{ x \in \mathbb{R} : d_0(x) < 0 \text{ and } \frac{1 + d_0(x)}{\rho_0(x)} + 2 \log \left(1 - \frac{d_0(x)}{2\rho_0(x)} \right) \leq 0 \right\}.$$

Proof Set $\beta = d/\rho$. Then it follows from (25) that β satisfies

$$\beta' = \frac{1}{\rho^2} (d'\rho - d\rho') = \frac{1}{\rho^2} (-d\rho + 2\rho^2 - \rho M_0) = -\beta + 2 - \frac{M_0}{\rho} \leq -\beta + 2.$$

This yields

$$\beta \leq 2 + (\beta_0 - 2)e^{-t} \quad \text{for } t \geq 0.$$

On the other hand, it follows from the continuity equation that

$$\rho' = -\rho^2\beta, \quad \text{i.e.,} \quad \rho^{-1} = \rho_0^{-1} + \int_0^t \beta(s) ds \leq \rho_0^{-1} + 2t + (\beta_0 - 2)(1 - e^{-t}).$$

Set $f(t) := \rho_0^{-1} + 2t + (\beta_0 - 2)(1 - e^{-t})$, then

$$f_0 = \rho_0^{-1} > 0 \quad \text{and} \quad \lim_{t \rightarrow +\infty} f(t) = \infty.$$

On the other hand, $f'(t) = 2 + (\beta_0 - 2)e^{-t}$; thus, if there exists a $t_* > 0$ such that $f'(t_*) = 0$ and $f(t_*) \leq 0$, then the density ρ is blowing up until this time $t_* > 0$. Note that $f'(t_*) = 0$ implies $e^{-t_*} = 2/(2 - \beta_0)$. This yields $\beta_0 < 0$, i.e., $d_0 < 0$ due to $e^{-t_*} \in (0, 1)$. Then for $d_0 < 0$ we get

$$f(t_*) = \rho_0^{-1} + \beta_0 + 2t_* = \rho_0^{-1} + \beta_0 + 2 \log\left(\frac{2 - \beta_0}{2}\right).$$

Hence, if there exists a x such that

$$d_0(x) < 0 \quad \text{and} \quad \rho_0^{-1}(x) + \beta_0(x) + 2 \log\left(\frac{2 - \beta_0(x)}{2}\right) \leq 0,$$

then the life-span T of the solution (ρ, u) should be finite. Furthermore, the time T satisfies

$$T \leq 2 \inf_{x \in \mathcal{S}} \log\left(\frac{2 - \beta_0(x)}{2}\right).$$

A study concerning the qualitative properties of the dynamics of the system (24) with the potential (23) is also conducted. Depending on the initial conditions, the density and position may converge to the steady state with or without oscillations around that state. In Fig. 16, it is depicted how the density and position of the boundary nodes evolve depending on the initial mass or the initial velocity. Generally, it shows in Fig. 16 (a) and (c) that there is a limit mass below in which there are no oscillations. However, this limit mass may change with the initial velocity. With respect to the influence of the initial velocity for a fixed mass, and according to Fig. 16 (b) and (d), it is possible to deduce that the more negative the initial slope of the velocity is, the larger the tendency to the oscillations will be.

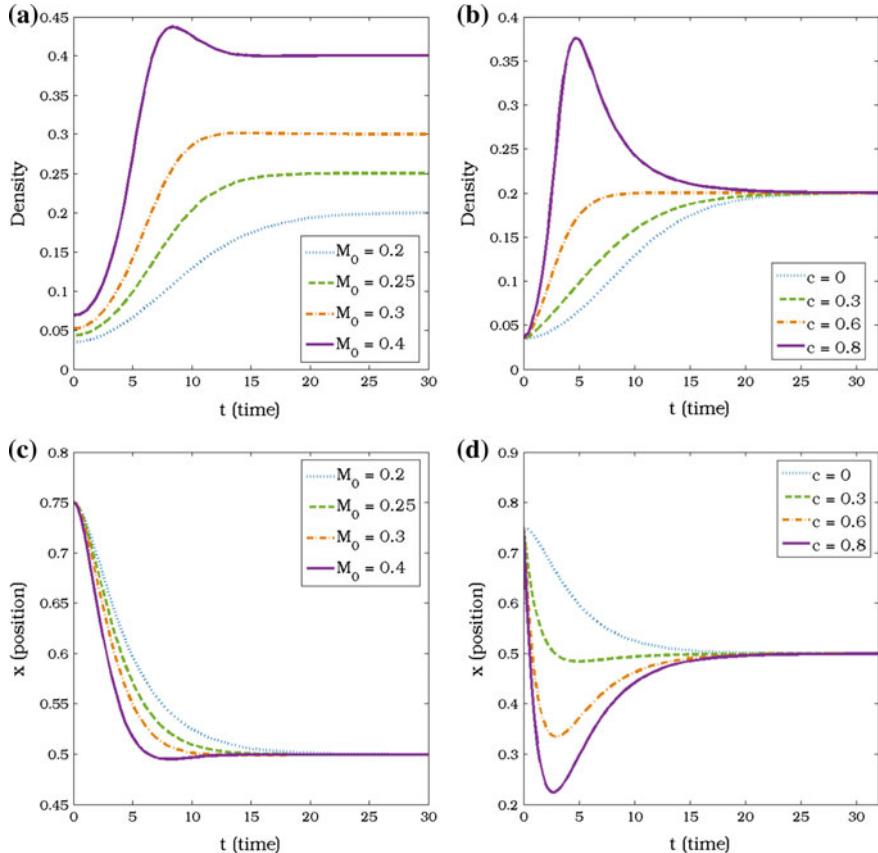


Fig. 16 Dynamical behavior of the convergence of boundary nodes. (a): Density with different values of the mass M_0 and $c = 0$. (b): Density with different values of the parameter c and $M_0 = 0.2$. (c): Position with different values of the mass M_0 and $c = 0$. (d): Position with different values of the parameter c and $M_0 = 0.2$.

Finally, some numerical simulations with the CS nonlocal velocity alignment instead of linear damping are conducted. Although no theoretical result is known for that case, our numerical simulations demonstrate that the total mass of the system and the initial conditions affect the global behavior of the solution in similar way as in the linear damping case.

In Fig. 17, two different simulations are depicted, and they differ in the initial velocity. The first one, Fig. 17 (a), corresponds to $M_0 = 1$ and $c = 0.2$, and we observe that there is no blowup in finite time reaching eventually the steady state. In contrast, the other simulation, increasing the value of the parameter c up to 0.5, shows finite-time blowup due to the infinite slope of the velocity on the boundary. The numerical time of blowup is $t = 2.22$.

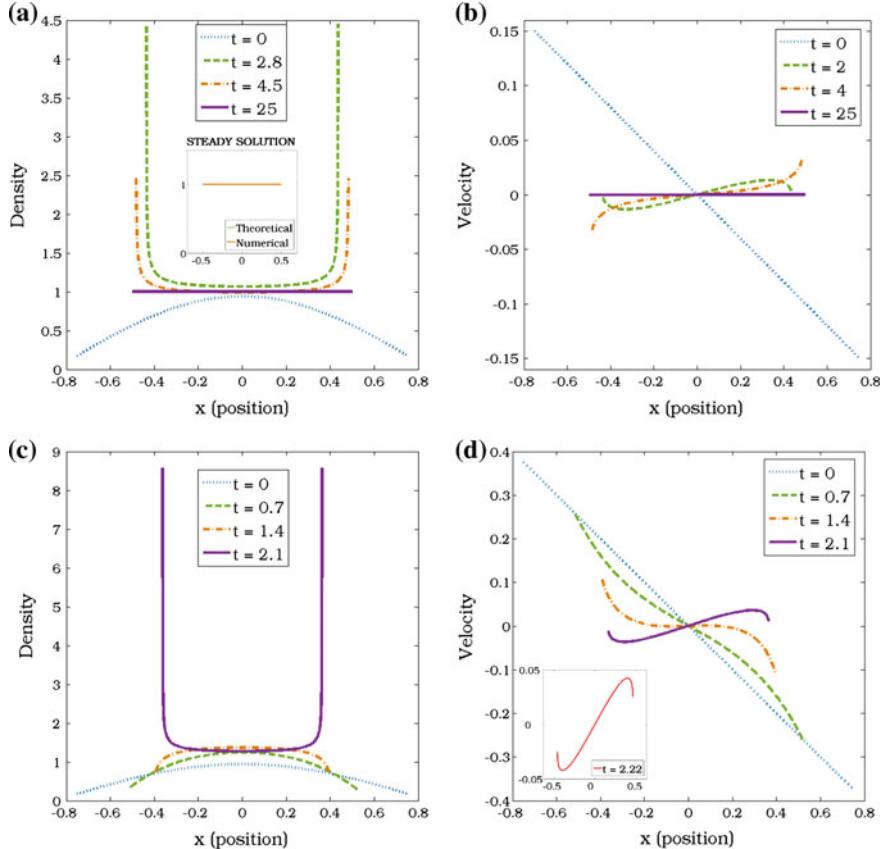


Fig. 17 Numerical simulations with CS nonlocal velocity alignment. (a), (b): Time behavior of the density and the velocity for the case $M_0 = 1$ and $c = 0.2$. (c), (d): Time behavior of the density and the velocity for the case $M_0 = 1$ and $c = 0.2$.

3.3.3 Stronger Repulsive Potential Than Newtonian

In this part, we deal with a potential more repulsive than Newtonian at the origin confined by the quadratic attractive potential given by

$$K(x) = -\log|x| + \frac{x^2}{2}. \quad (26)$$

For simplicity, we again consider the linear damping in the momentum equation. Steady solutions for this problem correspond again to find density profiles ρ satisfying (10) together with $u = 0$ on the support of the density ρ . They can be characterized as the global minimizer of certain energy functional and they can be computed explicitly, see [39] and [92]. In fact, they are given by the semicircle law; that is, their graph is a semicircle.

We choose as initial density for the numerical simulations the positive part of a cosine function. For each node $i \in \{1, \dots, N\}$, it is given by

$$\rho_i(0) = \frac{1}{\gamma} \cos\left(\pi \frac{x_i(0)}{1.5}\right),$$

where the constant γ is computed so that the total mass $M_0 := \int_{\mathbb{R}} \rho_0 dx$ has the required value. With respect to the initial velocity, it is chosen to be

$$u_i(0) = -c x_i(0) \quad \text{for } i = 1, \dots, n,$$

where the constant $c \in \mathbb{R}^+$ will be varied to study different initial conditions in the simulations.

We first find a blowup estimate for the system (24) with the potential (26). By differentiating the momentum equation of system (24), it is obtained that

$$\begin{aligned} \rho' &= -\rho d, \\ d' &= -\left(d^2 + d + \int_{\Omega(t)} \frac{\rho(y)}{|x-y|^2} dy + M_0\right) \leq -(d^2 + d + M_0), \end{aligned} \tag{27}$$

and for the case of $1 - 4M_0 \geq 0$ it is found that

$$\begin{aligned} \rho' &= -\rho d, \\ d' &\leq -(d - d_-)(d - d_+), \end{aligned} \tag{28}$$

where

$$d_{\pm} := \frac{-1 \pm \sqrt{1 - 4M_0}}{2}.$$

Theorem 6 *Let (ρ, u) be a classical solution to the system (24) with the potential (26) on the time interval $[0, T]$. Suppose that $1 - 4M_0 \geq 0$. Then the life span T of the solution (ρ, u) should be finite if there exists a x such that $\partial_x u_0(x) < \frac{-1 - \sqrt{1 - 4M_0}}{2}$. Moreover,*

$$T \leq \frac{1}{d_- - d_0}.$$

Proof We divide the proof into two steps.

Step 1. If $d_0 < d_-$, then $d(t) < d_-$ for $t \in [0, T]$. Set

$$\mathcal{T} := \{t \in [0, T] : d(s) < d_- \text{ for } s \in [0, t]\}.$$

Then \mathcal{T} is not empty since $0 \in \mathcal{T}$. Furthermore, if we set $\mathcal{T}^* = \sup \mathcal{T}$, then $\mathcal{T}^* > 0$ since $d(t)$ is continuous. Suppose that $\mathcal{T}^* < T$, then we get $\lim_{t \rightarrow \mathcal{T}^*-} d(t) = d_-$. On the other hand, it follows from (28) that

$$d' \leq -(d - d_-)(d - d_+) \quad \text{for } t \in [0, \mathcal{T}^*),$$

and this yields $d'(t) < 0$ for $t \in [0, \mathcal{T}^*)$ since $d(t) < d_- < d_+$ for $t \in [0, \mathcal{T}^*)$. This is a contradiction to

$$d_- = \lim_{t \rightarrow \mathcal{T}^*-} d(t) \leq d_0 < d_-.$$

Thus, $\mathcal{T}^* \geq T$ and $d(t) < d_-$ for $t \in [0, T]$.

Step 2. If $d_0 < d_-$, then the life span of smooth solutions T should be finite. Since $d(t) < d_- < d_+$ for $t \in [0, T]$, we get

$$d' \leq -(d - d_-)(d - d_+) \leq -(d - d_-)^2, \quad \text{i.e., } (d - d_-)' \leq -(d - d_-)^2.$$

This implies

$$d(t) \leq \frac{1}{(d_0 - d_-)^{-1} + t} + d_-.$$

Since $d_0 < d_-$, the right-hand side of the above equality diverges to $-\infty$ when $t \rightarrow (d_- - d_0)^{-1}$. This concludes that the life-span T should be less than $(d_- - d_0)^{-1}$.

In Fig. 18, two numerical simulations are depicted with the objective of supporting Theorem 6. In the first case, with $M_0 = 0.2$ and $c = 0.3$, the initial conditions of the blowup estimate of Theorem 6 are not satisfied. The numerical experiment seems to indicate that there exists a global-in-time solution converging toward the steady solution given by the semicircle law. We do not have sharp critical thresholds for this system so its behavior could not be predicted beforehand. On the other hand, in the second case it is set $M_0 = 0.2$ and $c = 1$, and those initial conditions satisfy the blowup estimate of Theorem 6. It can be observed in Fig. 18 (d) that the finite-time blowup is again produced by the infinite slope of the velocity at the boundary. Finding critical thresholds for this case is an open problem.

To conclude, a numerical study of the CS system (21) with (26) has been carried out. In fact, as above the total mass of the system and the initial conditions affect the global behavior of the solution leading to global existence of solutions converging to the semicircle law or finite-time blowup. The qualitative behavior is similar to the case of linear damping in case of initial data with zero mean velocity. Otherwise, the system can lead to traveling wave solutions with the same density profile.

3.3.4 Euler-Alignment System with Attractive Power Law Potential and Pressure

In this part, we consider the pressure term p in the Euler-alignment system with the attractive power law potential $K(x) = x^2/2$ leading to the system

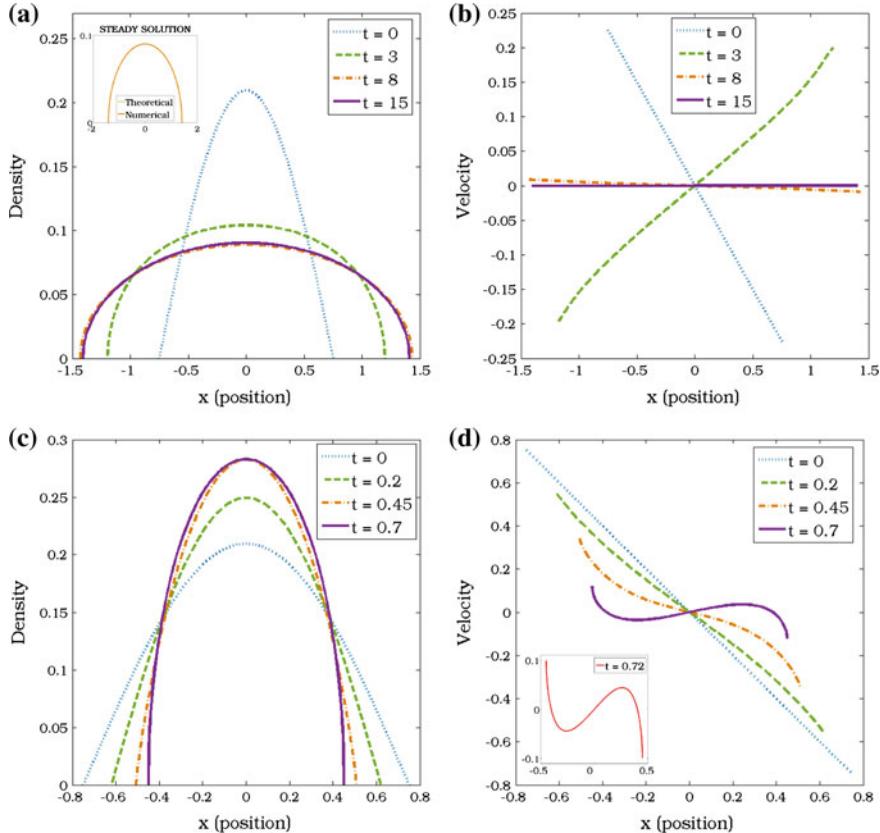


Fig. 18 Numerical simulations of the system (24) with potential (26). (a), (b): Time behavior of the density and the velocity for the case $M_0 = 0.2$ and $c = 0.3$. (c), (d): Time behavior of the density and the velocity for the case $M_0 = 0.2$ and $c = 1$.

$$\begin{cases} \partial_t \rho + \partial_x (\rho u) = 0, & x \in \mathbb{R}, \quad t > 0, \\ \partial_t (\rho u) + \partial_x (\rho u^2) + \partial_x p(\rho) = \rho [\psi * (\rho u) - (\psi * \rho)u - \partial_x K \star \rho], \end{cases} \quad (29)$$

where the pressure law is given by $p(\rho) = \rho^2$.

For the numerical approximation, we will use a variation of the Lagrangian scheme as in Section 3.1. Notice that we cannot directly apply that scheme due to the presence of pressure p . In order to overcome this new difficulty, we take into account the approximated pressure term $\partial_x p_\varepsilon(\rho) := 2\rho(\partial_x \delta_\varepsilon \star \rho)$, where δ_ε is a mollification of the Dirac delta function δ_0 given by a Gaussian

$$\delta_\varepsilon(x) = \frac{1}{\sqrt{2\pi\varepsilon}} e^{-\frac{x^2}{2\varepsilon}}.$$

Note that δ_ε converges weakly as measures to δ_0 as $\varepsilon \rightarrow 0$, and subsequently, this implies $\partial_x p_\varepsilon(\rho) = 2\rho(\delta_\varepsilon \star \partial_x \rho) \rightarrow 2\rho \partial_x \rho = \partial_x p(\rho)$ as $\varepsilon \rightarrow 0$, for smooth enough mass densities ρ . Using this approximation, we can rewrite the system (29)₂ as

$$\partial_t(\rho u) + \partial_x(\rho u^2) = \int_{\mathbb{R}} \psi(x-y)(u(y)-u(x))\rho(x)\rho(y) dy - \rho \left(\partial_x \tilde{K} \star \rho \right),$$

where the interaction potential \tilde{K} is given by

$$\tilde{K}(x) = -\frac{1}{\sqrt{2\pi\varepsilon}} e^{\frac{-x^2}{2\varepsilon}} + \frac{x^2}{2}.$$

This enables us to use the previous Lagrangian scheme for the numerical simulations. It is worth mentioning that the parameter ε cannot be too small for the numerical simulation and it should be chosen according to both the number of nodes and the distance between them. In our setting, we take the parameter $\varepsilon = 10^{-4.1}$ for the numerical simulations.

Similarly as before, we consider the CS nonlocal velocity alignment force and the linear damping for the numerical simulations. We again remind the reader that if we choose the constant communication function $\psi \equiv 1$, then CS velocity alignment force becomes the linear damping by assuming that the initial momentum is zero, i.e., $\int_{\mathbb{R}}(\rho_0 u_0)(x) dx = 0$.

Let us next investigate the steady solution for the system (29). Let us first look for steady solutions of the form $\rho = \rho_\infty(x)$ and $u = u_\infty \equiv 0$. Since the initial momentum is zero, it is straightforward to check that the center of mass of the density ρ is preserved on time. Let us assume without loss of generality that the center of mass is zero. Plugging ρ_∞ and u_∞ into (29), we easily find

$$2\partial_x \rho_\infty(x) = -(x \star \rho_\infty)(x) = -x M_0 \quad \text{on } \text{supp}(\rho_\infty).$$

This yields, by solving the ODE and fixing the mass to be M_0 with zero center of mass, that

$$\rho_\infty(x) = \begin{cases} -\frac{M_0}{4} (x + \sqrt[3]{3}) (x - \sqrt[3]{3}) & \text{for } x \in [-\sqrt[3]{3}, \sqrt[3]{3}], \\ 0 & \text{otherwise.} \end{cases}$$

The initial density and velocity for the numerical simulations are defined as

$$\rho_i(0) = \frac{1}{\gamma} \cos \left(\pi \frac{x_i(0)}{1.5} \right) \quad \text{and} \quad u_i(0) = -c \sin \left(\pi \frac{x_i(0)}{1.5} \right),$$

for each node $i = 1, \dots, n$, where γ is chosen so that the mass of the system is unit, and $c = 0.2$. Actually, in our numerical experiments we add to the initial data the positive constant $\alpha = 0.05$. The presence of vacuum areas leads to numerical artifacts

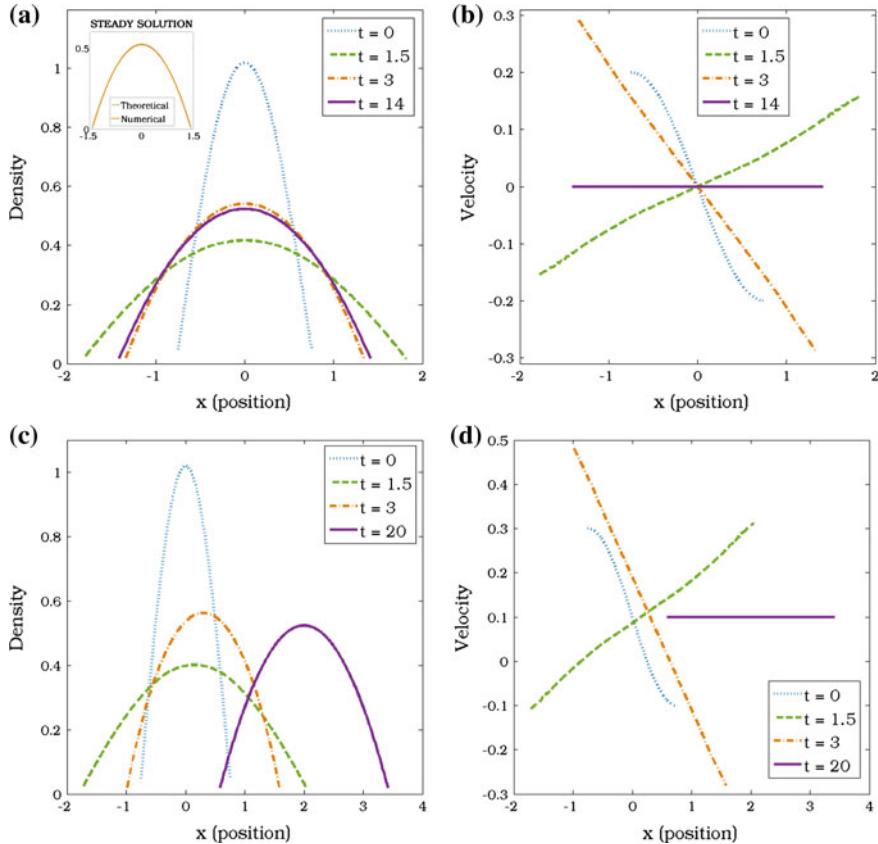


Fig. 19 Numerical simulation of the density and velocity with the linear damping (a), (b) and the CS velocity alignment (c), (d).

at the boundary if we strictly impose zero value of the density. This is mainly due to the pressure term since the density is expected to become non-differentiable at the tip of the support, as for the steady-state ρ_∞ . Therefore, we opt by adding this small constant to our initial data over the whole simulation interval. Furthermore, for the CS nonlocal velocity alignment force we have added a constant value to the initial velocity of 0.1, in order to have a nonzero initial momentum.

Fig. 19 shows the time evolutions of the density and the velocity for the approximated system with the linear damping (a), (b) and the CS nonlocal velocity alignment force (c), (d). We observe a convergence to the steady state for linear damping, while for the CS model a convergence toward a traveling wave profile due to the nonzero initial momentum. In both cases, we observe very fast convergences toward the steady-state/traveling wave with time-modulated decaying oscillations. Furthermore, it shows that the shape of the asymptotic density profiles is consistent with the theoretical results.

Acknowledgements J. A. C. was partially supported by the Royal Society via a Wolfson Research Merit Award. J. A. C. and Y. -P. C. were partially supported by EPSRC grant EP/K008404/1. Y. -P. C. was supported by the ERC-Stating grant HDSPCONTR “High-Dimensional Sparse Optimal Control”. S. P. P. was partially supported by a Erasmus+ scholarship.

References

1. S. Ahn, H. Choi, S.-Y. Ha, and H. Lee, *On the collision avoiding initial-configurations to the Cucker-Smale type flocking models*, *Comm. Math. Sci.*, 10:625–643, 2012.
2. M. Aguech, R. Illner, and A. Richardson, *Analysis and simulations of a refined flocking and swarming model of Cucker-Smale type*, *Kinetic and Related Models* 4:1–16, 2011.
3. G. Albi, D. Balagué, J. A. Carrillo, J. von Brecht, *Stability analysis of flock and mill rings for second order models in swarming*, *SIAM J. Appl. Math.*, 74:794–818, 2014.
4. G. Albi, L. Pareschi, *Modelling self-organized systems interacting with few individuals: from microscopic to macroscopic dynamics*, *Applied Math. Letters*, 26:397–401, 2013.
5. I. Aoki, *A Simulation Study on the Schooling Mechanism in Fish*, *Bull. Jap. Soc. Sci. Fisheries* 48:1081–1088, 1982.
6. H.-O. Bae, Y.-P. Choi, S.-Y. Ha, and M.-J. Kang, *Time-asymptotic interaction of flocking particles and incompressible viscous fluid*, *Nonlinearity* 25:1155–1177, 2012.
7. H.-O. Bae, Y.-P. Choi, S.-Y. Ha, and M.-J. Kang, *Asymptotic flocking dynamics of Cucker-Smale particles immersed in compressible fluids*, *Disc. and Cont. Dyn. Sys.* 34:4419–4458, 2014.
8. H.-O. Bae, Y.-P. Choi, S.-Y. Ha, and M.-J. Kang, *Global existence of strong solution for the Cucker-Smale-Navier-Stokes system*, *J. Diff. Eqns.* 257:2225–2255, 2014.
9. H.-O. Bae, Y.-P. Choi, S.-Y. Ha, and M.-J. Kang, *Global existence of strong solutions to the Cucker-Smale-Stokes system*, *J. Math. Fluid Mech.* 18:381–396, 2016.
10. D. Balagué, and J. A. Carrillo, *Aggregation equation with growing at infinity attractive-repulsive potentials*, *Proceedings of the 13th International Conference on Hyperbolic Problems, Series in Contemporary Applied Mathematics CAM 17*, Higher Education Press, 1:136–147, 2012.
11. D. Balagué, Carrillo, T. J. A., Laurent, and G. Raoul, *Nonlocal interactions by repulsive-attractive potentials: radial ins/stability*, *Physica D*, 260:5–25, 2013.
12. D. Balagué, Carrillo, T. J. A., Laurent, and G. Raoul, *Dimensionality of Local Minimizers of the Interaction Energy*, *Arch. Rat. Mech. Anal.*, 209:1055–1088, 2013.
13. A. Barbaro, J. A. Cañizo, J. A. Carrillo, P. Degond, *Phase Transitions in a kinetic flocking model of Cucker-Smale type*, *Multiscale Model. Simul.* 14:1063–1088, 2016.
14. A. 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. and Computers in Simulation*, 79:3397–3414, 2009.
15. N. Bellomo, C. Dogbe, *On the modeling of traffic and crowds: A survey of models, speculations, and perspectives*, *SIAM Review* 53:409–463, 2011.
16. A. J. Bernoff, C. M. Topaz, *A primer of swarm equilibria*, *SIAM J. Appl. Dyn. Syst.*, 10:212–250, 2011.
17. A. L. Bertozzi, J. A. Carrillo, and T. Laurent, *Blowup in multidimensional aggregation equations with mildly singular interaction kernels*, *Nonlinearity*, 22:683–710, 2009.
18. A. L. Bertozzi, T. Kolokolnikov, H. Sun, D. Uminsky, J. von Brecht, *Ring patterns and their bifurcations in a nonlocal model of biological swarms*, *Commun. Math. Sci.*, 13:955–985, 2015.
19. A. L. Bertozzi and T. Laurent, *Finite-time blow-up of solutions of an aggregation equation in \mathbb{R}^n* , *Comm. Math. Phys.*, 274:717–735, 2007.
20. A. L. Bertozzi, T. Laurent, and J. Rosado, *L^p theory for the multidimensional aggregation equation*, *Comm. Pure Appl. Math.*, 43:415–430, 2010.

21. A. L. Bertozzi, T. Laurent, and F. Léger, *Aggregation and spreading via the newtonian potential: the dynamics of patch solutions*, *Mathematical Models and Methods in Applied Sciences*, 22(supp01):1140005, 2012.
22. M. Bodnar, J.J.L. Velazquez, *Friction dominated dynamics of interacting particles locally close to a crystallographic lattice*, *Math. Methods Appl. Sci.*, 36:1206–1228, 2013.
23. F. Bolley, J. A. Cañizo, and J. A. Carrillo *Stochastic mean-field limit: non-Lipschitz forces & swarming*, *Math. Mod. Meth. Appl. Sci.*, 21:2179–2210, 2011.
24. M. Bostan, J. A. Carrillo, *Asymptotic fixed-speed reduced dynamics for kinetic equations in swarming*, *Math. Models Methods Appl. Sci.* 23:2353–2393, 2013.
25. W. Braun and K. Hepp, *The Vlasov Dynamics and Its Fluctuations in the 1/N Limit of Interacting Classical Particles*, *Commun. Math. Phys.*, 56:101–113, 1977.
26. M. Burger, P. Markowich, and J. Pietschmann, *Continuous limit of a crowd motion and herding model: Analysis and numerical simulations*, *Kinetic and Related Methods*, 4:1025–1047, 2011.
27. S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau, *Self-Organization in Biological Systems*, Princeton University Press, 2003.
28. J.A. Cañizo, J.A. Carrillo, and J. Rosado, *Collective Behavior of Animals: Swarming and Complex Patterns*, *Arbor*, 186:1035–1049, 2010.
29. J.A. Cañizo, J.A. Carrillo, and J. Rosado, *A well-posedness theory in measures for some kinetic models of collective motion*, *Math. Mod. Meth. Appl. Sci.*, 21:515–539, 2011.
30. J. A. Carrillo, Y.-P. Choi, and M. Hauray, *The derivation of swarming models: Mean-field limit and Wasserstein distances*, *Collective Dynamics from Bacteria to Crowds: An Excursion Through Modeling, Analysis and Simulation, Series: CISM International Centre for Mechanical Sciences*, Springer, 533:1–45, 2014.
31. J. A. Carrillo, Y.-P. Choi, and M. Hauray, *Local well-posedness of the generalized Cucker-Smale model with singular kernels*, *ESAIM Proc.*, 47:17–35, 2014.
32. J. A. Carrillo, Y.-P. Choi, M. Hauray, and S. Salem, *Mean-field limit for collective behavior models with sharp sensitivity regions*, to appear in *J. Eur. Math. Soc.*.
33. J. A. Carrillo, Y.-P. Choi, and T. Karper, *On the analysis of a coupled kinetic-fluid model with local alignment forces*, *Annales de l'IHP-ANL*, 33:273–307, 2016.
34. J. A. Carrillo, Y.-P. Choi, E. Tadmor, and C. Tan, *Critical thresholds in 1D Euler equations with nonlocal forces*, *Math. Mod. Meth. Appl. Sci.*, 26:185–206, 2016.
35. J. A. Carrillo, Y.-P. Choi, and E. Zatorska, *On the pressureless damped Euler-Poisson equations with quadratic confinement: Critical thresholds and large-time behavior*, *Math. Models Methods Appl. Sci.* 26:2311–2340, 2016.
36. J. A. Carrillo, M. Di Francesco, A. Figalli, T. Laurent, and D. Slepčev, *Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations*, *Duke Math. J.*, 156:229–271, 2011.
37. J. A. Carrillo, M. Di Francesco, A. Figalli, T. Laurent, and D. Slepčev, *Confinement in nonlocal interaction equations*, *Nonlinear Anal.*, 75(2):550–558, 2012.
38. J. A. Carrillo, M. R. D’Orsogna, and V. Panferov, *Double milling in self-propelled swarms from kinetic theory*, *Kinetic and Related Models* 2:363–378, 2009.
39. J.A. Carrillo, L.C.F. Ferreira, J.C. Precioso, *A mass-transportation approach to a one dimensional fluid mechanics model with nonlocal velocity*, *Advances in Mathematics*, 231:306–327, 2012.
40. 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.
41. J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil, *Particle, Kinetic, and Hydrodynamic Models of Swarming*, *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences, Series: Modelling and Simulation in Science and Technology*, Birkhäuser, 297–336, 2010.
42. J. A. Carrillo, Y. Huang, *Explicit Equilibrium Solutions For the Aggregation Equation with Power-Law Potentials*, *Kinetic Rel. Mod.* 10:171–192, 2017.
43. J. A. Carrillo, Y. Huang, S. Martin, *Explicit flock solutions for Quasi-Morse potentials*, *European J. Appl. Math.*, 25:553–578, 2014.

44. J. A. Carrillo, Y. Huang, S. Martin, *Nonlinear stability of flock solutions in second-order swarming models*, *Nonlinear Anal. Real World Appl.*, 17:332–343, 2014.
45. J. A. Carrillo, A. Klar, S. Martin, and S. Tiwari, *Self-propelled interacting particle systems with roosting force*, *Math. Mod. Meth. Appl. Sci.*, 20:1533–1552, 2010.
46. J. A. Carrillo, A. Klar, A. Roth, *Single to double mill small noise transition via semi-lagrangian finite volume methods*, *Comm. Math. Sci.* 14:1111–1136, 2016.
47. J. A. Carrillo, S. Martin, V. Panferov, *A new interaction potential for swarming models*, *Physica D*, 260:112–126, 2013.
48. J.A. Carrillo, R.J. McCann, and C. Villani, *Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates*, *Rev. Matemática Iberoamericana*, 19:1–48, 2003.
49. J.A. Carrillo, R.J. McCann, and C. Villani, *Contractions in the 2-Wasserstein length space and thermalization of granular media*, *Arch. Rat. Mech. Anal.*, 179:217–263, 2006.
50. Y.-P. Choi, *Global classical solutions of the Vlasov-Fokker-Planck equation with local alignment forces*, *Nonlinearity*, 29:1887–1916, 2016.
51. Y.-P. Choi, *Compressible Euler equations intreating with incompressible flow*, *Kinetic and Related Models*, 8:335–358, 2015.
52. Y.-L. Chuang, M. R. D’Orsogna, D. Marthaler, A. L. Bertozzi, L. S. Chayes, *State transitions and the continuum limit for a 2D interacting self-propelled particle system*, *Phys. D* 232:33–47, 2007.
53. I. D. Couzin, J. Krause, *Self-organization and collective behavior of vertebrates*, *Adv. Study Behav.* 32:1–67, 2003.
54. F. Cucker and S. Smale, *On the mathematics of emergence*, *Japan. J. Math.* 2:197–227, 2007.
55. F. Cucker and S. Smale, *Emergent behavior in flocks*, *IEEE Trans. Automat. Control* 52:852–862, 2007.
56. P. Degond, A. Frouvelle, J.-G. Liu, *Macroscopic limits and phase transition in a system of self-propelled particles*, *J. Nonlinear Sci.* 23:427–456, 2013.
57. P. Degond, A. Frouvelle, J.-G. Liu, *Phase transitions, hysteresis, and hyperbolicity for self-organized alignment dynamics*, *Arch. Ration. Mech. Anal.* 216:63–115, 2015.
58. P. Degond, S. Motsch, *Continuum limit of self-driven particles with orientation interaction*, *Math. Models Methods Appl. Sci* 18 supp01:1193–1215, 2008.
59. R. Dobrushin, *Vlasov equations*, *Funct. Anal. Appl.* 13:115–123, 1979.
60. M. R. D’Orsogna, Y. L. Chuang, A. L. Bertozzi, and L. Chayes, *Self-propelled particles with soft-core interactions: patterns, stability, and collapse*, *Phys. Rev. Lett.* 96, 2006.
61. R. Duan, M. Fornasier, and G. Toscani, *A kinetic flocking model with diffusion*, *Comm. Math. Phys.*, 200:95–145, 2010.
62. R. C. Fetecau, Y. Huang, T. Kolokolnikov, *Swarm dynamics and equilibria for a nonlocal aggregation model*, *Nonlinearity*, 24:2681–2716, 2011.
63. N. Fournier, M. Hauray, and S. Mischler, *Propagation of chaos for the 2D viscous vortex model*, *J. Eur. Math. Soc.*, 16:1423–1466, 2014.
64. F. Golse, *The Mean-Field Limit for the Dynamics of Large Particle Systems*, *Journées équations aux dérivées partielles*, 9:1–47, 2003.
65. S.-Y. Ha, J.-G. Liu, *A simple proof of the Cucker-Smale flocking dynamics and mean-field limit*, *Commun. Math. Sci.* 7 (2) (2009) 297–325.
66. S.-Y. Ha and E. Tadmor, *From particle to kinetic and hydrodynamic descriptions of flocking*, *Kinetic and Related Models* 1:415–435, 2008.
67. J. Haskovec, *Flocking dynamics and mean-field limit in the Cucker-Smale-type model with topological interactions*, *Physica D*, 261:42–51, 2013.
68. M. Hauray, *Wasserstein distances for vortices approximation of Euler-type equations*, *Math. Mod. Meth. Appl. Sci.*, 19:1357–1384, 2009.
69. M. Hauray and P.-E. Jabin, *Particles approximations of Vlasov equations with singular forces: Propagation of chaos*, *Ann. Sci. Ec. Norm. Super.*, 48:891–940, 2015.
70. C. K. Hemelrijk and H. Hildenbrandt, *Self- Organized Shape and Frontal Density of Fish Schools*, *Ethology* 114, 2008.

71. H. Hildenbrandt, C. Carere, C. K. Hemelrijk, *Self-organized aerial displays of thousands of starlings: a model*, *Behavioral Ecology* 21:1349–1359, 2010.
72. A. Huth and C. Wissel, *The Simulation of the Movement of Fish Schools*, *J. Theo. Bio.*, 1992.
73. Y. Katz, K. Tunstrom, C. C. Ioannou, C. Huepe, I. D. Couzin, *Inferring the structure and dynamics of interactions in schooling fish*, *PNAS*, 108:18720–18725, 2011.
74. A. Klar and S. Tiwari, *A multiscale meshfree method for macroscopic approximations of interacting particle systems*, *Multiscale Model. Simul.*, 12:1167–1192, 2014.
75. T. Kolokolnikov, J. A. Carrillo, A. Bertozzi, R. Fetecau, M. Lewis, *Emergent behaviour in multi-particle systems with non-local interactions*, *Phys. D*, 260:1–4, 2013.
76. T. Kolokonikov, H. Sun, D. Uminsky, and A. Bertozzi, *Stability of ring patterns arising from 2d particle interactions*, *Physical Review E*, 84:015203, 2011.
77. C. Lattanzio, A. E. Tzavaras, *Relative entropy in diffusive relaxation*, *SIAM J. Math. Anal.* 45:1563–1584, 2013.
78. T. Laurent, *Local and global existence for an aggregation equation*, *Communications in Partial Differential Equations*, 32:1941–1964, 2007.
79. H. Levine, W.-J. Rappel and I. Cohen, *Self-organization in systems of self-propelled particles*, *Phys. Rev. E*, 63:017101, 2000.
80. A. J. Leverentz, C. M. Topaz, A. J. Bernoff, *Asymptotic dynamics of attractive-repulsive swarms*, *SIAM J. Appl. Dyn. Syst.*, 8:880–908, 2009.
81. Y. X. Li, R. Lukeman, and L. Edelstein-Keshet, *Minimal mechanisms for school formation in self-propelled particles*, *Physica D*, 237:699–720, 2008.
82. R. Lukeman R, Y. X. Li, L. Edelstein-Keshet, *How do ducks line up in rows: inferring individual rules from collective behaviour*, *PNAS*, 107:12576–12580, 2010.
83. A. Mogilner, L. Edelstein-Keshet, L. Bent, and A. Spiros, *Mutual interactions, potentials, and individual distance in a social aggregation*, *J. Math. Biol.*, 47:353–389, 2003.
84. A. Mogilner, L. Edelstein-Keshet, *A non-local model for a swarm*, *J. Math. Bio.*, 38:534–570, 1999.
85. S. Motsch, E. Tadmor, *A new model for self-organized dynamics and its flocking behavior*, *J. Stat. Phys.*, 144:923–947, 2011.
86. S. Motsch, E. Tadmor, *Heterophilious dynamics enhances consensus*, *SIAM Review* 56:577–621, 2014.
87. H. Neunzert, *An introduction to the nonlinear Boltzmann–Vlasov equation*, In *Kinetic theories and the Boltzmann equation (Montecatini Terme, 1981)*, *Lecture Notes in Math.* 1048. Springer, Berlin, 1984.
88. K. J. Painter, J. M. Bloomfield, J. A. Sherratt, A. Gerisch, *A nonlocal model for contact attraction and repulsion in heterogeneous populations*, *Bulletin of Mathematical Biology* 77:1132–1165, 2015.
89. J. Parrish, and L. Edelstein-Keshet, *Complexity, pattern, and evolutionary trade-offs in animal aggregation*, *Science*, 294: 99–101, 1999.
90. M.J.D. Powell, *A Fortran Subroutine for Solving Systems of Nonlinear Algebraic Equations*, *Numerical Methods for Nonlinear Algebraic Equations*, (P. Rabinowitz, ed.), Ch.7, 1970.
91. C. W. Reynolds, *Flocks, herds and schools: A distributed behavioral model*, *ACM SIGGRAPH Computer Graphics*, 21: 25–34, 1987.
92. E. B. Saff and V. Totik, *Logarithmic Potentials with External Fields*, Springer, Berlin, 1997.
93. H. Spohn, *Large scale dynamics of interacting particles*, *Texts and Monographs in Physics*, Springer, 1991.
94. A.-S. Sznitman, *Topics in propagation of chaos*, In *Ecole d'Eté de Probabilités de Saint-Flour XIX 1989*, *Lecture Notes in Math.* 1464. Springer, Berlin, 1991.
95. E. Tadmor and C. Tan, *Critical thresholds in flocking hydrodynamics with non-local alignment*, *Phil. Trans. R. Soc. A*, 372:20130401, 2014.
96. C. Tan, *A discontinuous Galerkin method on kinetic flocking models*, to appear in *Math. Models Methods Appl. Sci.*
97. C.M. Topaz and A.L. Bertozzi, *Swarming patterns in a two-dimensional kinematic model for biological groups*, *SIAM J. Appl. Math.*, 65:152–174, 2004.

98. C.M. Topaz, A.L. Bertozzi, and M.A. Lewis, *A nonlocal continuum model for biological aggregation*, *Bulletin of Mathematical Biology*, 68:1601–1623, 2006.
99. T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Shochet, *Novel type of phase transition in a system of self-driven particles*, *Phys. Rev. Lett.*, 75:1226–1229, 1995.

Emergent Dynamics of the Cucker–Smale Flocking Model and Its Variants

Young-Pil Choi, Seung-Yeal Ha and Zhuchun Li

Abstract In this chapter, we present the Cucker–Smale-type flocking models and discuss their mathematical structures and flocking theorems in terms of coupling strength, interaction topologies, and initial data. In 2007, two mathematicians Felipe Cucker and Steve Smale introduced a second-order particle model which resembles Newton’s equations in N -body system and present how their simple model can exhibit emergent flocking behavior under sufficient conditions expressed only in terms of parameters and initial data. After Cucker–Smale’s seminal works in [31, 32], their model has received lots of attention from applied math and control engineering communities. We discuss the state of the art for the flocking theorems to Cucker–Smale-type flocking models.

1 Introduction

The jargon “*flocking*” represents collective phenomena in which self-propelled particles (or agents) are organized into an ordered motion from a disordered state using only limited environmental information and simple rules [63]. Such an organized motion is ubiquitous in our nature, e.g., aggregation of bacteria, flocking of birds, swarming of fish, and herding of sheep [6, 64], and they have been extensively studied recently because of their possible applications to sensor networks, controls of

Y.-P. Choi

Fakultät für Mathematik, Technische Universität München, Boltzmannstraße 3,
85748 Garching bei München, Germany
e-mail: ychoi@ma.tum.de

S.-Y. Ha (✉)

Department of Mathematical Sciences and Research Institute of Mathematics,
Seoul National University, Seoul 151-747, Republic of Korea
e-mail: syha@snu.ac.kr

Z. Li

Department of Mathematics, Harbin Institute of Technology, Harbin 150001,
People’s Republic of China
e-mail: lizhuchun@hit.edu.cn

robots, and unmanned aerial vehicles [49, 55, 57], and opinion formation of social networks. After the pioneering work [58, 64] of Reynolds and Vicsek et al, many agent-based models have been proposed in the literature and studied extensively both analytically and numerically. Among them, we are interested in the model introduced by Cucker and Smale [31, 32]. This model resembles a Newton-type N -body system for an interacting particle system. In the sequel, we introduce Cucker–Smale (C-S)-type models from microscopic to macroscopic scales and under various network topologies. We also summarize the state-of-the-art flocking theorems for the C-S-type models and explain how these models can achieve asymptotic flocking under what conditions and main ideas behind them. For other surveys on the related topics, we refer to [14, 51].

The rest of this chapter is organized as follows. In Section 2, we present hierarchical models for the description of C-S flocking ensemble starting from the particle to kinetic and fluid descriptions. In Section 3, we introduce three continuous-time C-S-type models including the original flocking model [32] and discuss the flocking problem for these models. In Section 4, we present discrete-time C-S models with leadership structures such as hierarchical and rooted leaders, alternating leadership. In Section 5, we present a mesoscopic description, namely kinetic picture for the C-S flocking. We also discuss flocking particle–fluid interactions via the coupled kinetic–fluid model. In Section 6, we present a C-S hydrodynamic flocking model and its flocking estimate, and then, we study its coupling with compressible Navier–Stokes equations through the drag force.

Notation: Throughout the chapter, we use a superscript to denote the component of a vector; for example, $x := (x^1, \dots, x^d) \in \mathbb{R}^d$. Subscripts are used to represent the ordering of particles. For vectors $x, v \in \mathbb{R}^d$, its ℓ_2 -norm and the inner product are defined as follows:

$$|x| := \left(\sum_{i=1}^d (x^i)^2 \right)^{\frac{1}{2}}, \quad \langle x, v \rangle := \sum_{i=1}^d x^i v^i,$$

where x^i and v^i are the i th components of x and v , respectively.

2 Preliminaries

In this section, we briefly discuss C-S models from microscopic to mesoscopic and macroscopic scales following the presentation in [42].

In [31, 32], Cucker and Smale introduced a Newton-type microscopic model for an interacting many-body system exhibiting a flocking phenomenon and provide sufficient conditions for an asymptotic flocking (see Definition 1) in terms of initial configuration and interaction topologies. We next describe the C-S model. Let x_i and v_i be the position and velocity of the i th C-S particle, respectively. Then, the C-S model with metric-dependent communication weight ψ is given by the following ODE system:

$$\begin{aligned}\frac{dx_i}{dt} &= v_i, \quad t > 0, \quad i = 1, \dots, N, \\ \frac{dv_i}{dt} &= \frac{K}{N} \sum_{j=1}^N \psi(|x_j - x_i|)(v_j - v_i),\end{aligned}\tag{1}$$

where K is a nonnegative coupling strength and ψ is a communication weight measuring the degree of communications (interactions) between particles. For a large C-S system (1) with $N \gg 1$, it is not reasonable to integrate the particle model for computational purpose, because it is too expensive to integrate (1) numerically even if it is possible. Thus, it is natural to introduce a kinetic model as an approximation for (1). For this, we introduce a kinetic density (one-particle distribution function) $f = f(x, \xi, t)$ at phase space (x, ξ) , at time t . Then, the spatial–temporal evolution of f is governed by the following Vlasov–McKean equation:

$$\begin{aligned}\partial_t f + \xi \cdot \nabla_x f + \nabla_\xi \cdot (F_a(f)f) &= 0, \quad (x, \xi) \in \mathbb{R}^d \times \mathbb{R}^d, \quad t > 0, \\ F_a(f)(x, \xi, t) &= -K \int_{\mathbb{R}^{2d}} \psi(|x - y|)(\xi - \xi_*)f(y, \xi_*, t)d\xi_*dy.\end{aligned}\tag{2}$$

The equation (2) admits a global smooth solution, as long as initial datum is compactly supported in x and v and sufficiently regular (see [42]). In kinetic theory of gases, it is well known that the velocity moments of f yield the macroscopic observables. For example, for a given $(x, t) \in \mathbb{R}^d \times \mathbb{R}_+$, we set

$$\begin{aligned}\rho &:= \int_{\mathbb{R}^d} f d\xi : \text{ local mass density,} \\ \rho u &:= \int_{\mathbb{R}^d} \xi f d\xi : \text{ local momentum density,} \\ \rho E &:= \rho e + \frac{1}{2} \rho |u|^2 : \text{ local energy density,}\end{aligned}\tag{3}$$

where $\rho e := \frac{1}{2} \int_{\mathbb{R}^d} |\xi - u(x)|^2 f d\xi$ is the internal energy. Then, macroscopic observables (3) satisfy the following hydrodynamic equations:

$$\begin{aligned}\partial_t \rho + \nabla_x \cdot (\rho u) &= 0, \quad x \in \mathbb{R}^d, \quad t > 0, \\ \partial_t(\rho u) + \nabla_x \cdot (\rho u \otimes u + P) &= S^{(1)}, \\ \partial_t(\rho E) + \nabla_x \cdot (\rho Eu + Pu + q) &= S^{(2)},\end{aligned}\tag{4}$$

where $P = (p_{ij})$ and $q = (q_1, \dots, q_d)$ are stress tensor and heat flow, respectively.

$$p_{ij} := \int_{\mathbb{R}^d} (\xi_i - u_i)(\xi_j - u_j) f d\xi, \quad q_i := \int_{\mathbb{R}^d} (\xi_i - u_i) |\xi - u|^2 f d\xi,\tag{5}$$

and the source terms are given by the following relations:

$$\begin{aligned} S^{(1)} &:= -K \int_{\mathbb{R}^d} \psi(|x-y|)(u(x) - u(y))\rho(x)\rho(y)dy, \\ S^{(2)} &:= -K \int_{\mathbb{R}^d} \psi(|x-y|)(E(x) + E(y) - u(x) \cdot u(y))\rho(x)\rho(y)dy. \end{aligned} \quad (6)$$

Of course, the moment system (4) is not closed as it is, because we need to know the third velocity moment of f to calculate the heat flux q in (5). So far, suitable closure conditions for (4) (e.g., the local Maxwellian for the Boltzmann equation) are not known. In a quasi-flocking regime, we may employ the mono-kinetic ansatz for f :

$$f(x, \xi, t) = \rho(x, t)\delta(\xi - u(x, t)), \quad x, \xi \in \mathbb{R}^d, \quad t > 0. \quad (7)$$

Then, under this mono-kinetic assumption (7), the stress tensor $P = (P_{ij})$ and heat flux q become zero:

$$p_{ij} = 0, \quad q_i = 0, \quad 1 \leq i, j \leq d.$$

Thus, in the quasi-flocking regime, the system (4)–(6) is reduced to the pressureless Euler system with a flocking dissipation:

$$\begin{aligned} \partial_t \rho + \nabla_x \cdot (\rho u) &= 0, \quad x \in \mathbb{R}^d, \quad t > 0, \\ \partial_t(\rho u) + \nabla_x \cdot (\rho u \otimes u) &= -K \rho \int_{\mathbb{R}^d} \psi(|x-y|)\rho(y)(u(x) - u(y))dy, \end{aligned} \quad (8)$$

Note that the energy equation in (4) can be derived from the equations for ρ and ρu , and the condition (7) will be valid only for the collisionless regime. However, when particles with different microscopic velocities collide, the mono-kinetic ansatz (7) will break down. Therefore, our system (8) should be regarded as a quasi-equilibrium model for the hydrodynamic description of the C-S ensemble.

Remark 1 If we consider other strong interaction forces such as local alignment and noise, then the density function f is close to a thermodynamical equilibrium $f \sim C_0 \rho e^{-|u-\xi|^2/2}$, and in this case, the dynamics can be well approximated by a compressible isothermal Euler equations with the velocity alignment force. This rigorous derivation is obtained in [45] by employing a relative entropy argument.

3 Continuous-time Cucker–Smale-Type Models

In this section, we discuss continuous-time C-S models and their flocking estimates. As discussed in the previous section, after Cucker–Smale’s seminal works in [31, 32], several variants of the C-S model have been introduced for better modelings including local and nonsymmetric interactions, collision avoidance, and formation

control in [52, 56]. In the following, we explain how flocking estimates for particle models can be obtained. We first consider a Cauchy problem for the C-S model:

$$\begin{aligned}\dot{x}_i &= v_i, \quad t > 0, \quad i = 1, \dots, N, \\ \dot{v}_i &= \frac{K}{N} \sum_{j=1}^N \psi(|x_j - x_i|)(v_j - v_i),\end{aligned}\tag{9}$$

subject to initial data

$$(x_i, v_i)(0) = (x_{i0}, v_{i0}),\tag{10}$$

where the communication weight function $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}$ is assumed to be Lipschitz continuous, nonnegative, and nonincreasing:

$$\psi \in \text{Lip}(\mathbb{R}_+; \mathbb{R}), \quad \psi \geq 0, \quad (\psi(r_2) - \psi(r_1))(r_2 - r_1) \geq 0, \quad r_1, r_2 \geq 0.\tag{11}$$

Before we present flocking estimates for (9)–(10), we recall the definition of (mono-cluster) flocking of a many-body system as follows.

Definition 1 [31, 42] Let $\mathcal{G} := \{(x_i, v_i)\}_{i=1}^N$ be an N -body interacting system. Then, \mathcal{G} exhibits a asymptotic flocking if and only if the following two relations hold.

1. (Velocity alignment): The relative velocities approach to zero asymptotically.

$$\lim_{t \rightarrow \infty} |v_i(t) - v_j(t)| = 0, \quad 1 \leq i, j \leq N.$$

2. (Spatial coherence): The relative positions are uniformly bounded:

$$\sup_{0 \leq t < \infty} |x_i(t) - x_j(t)| < \infty, \quad 1 \leq i, j \leq N.$$

To have some feeling for the large-time dynamics of (9), we consider the simplest system made of two C-S particles on the real line \mathbb{R} :

$$\begin{aligned}\dot{x}_1 &= v_1, \quad \dot{x}_2 = v_2, \quad t > 0, \quad x_i, v_i \in \mathbb{R}, \\ \dot{v}_1 &= \frac{K}{2} \psi(|x_2 - x_1|)(v_2 - v_1), \quad \dot{v}_2 = \frac{K}{2} \psi(|x_1 - x_2|)(v_1 - v_2), \\ (x_i, v_i)(0) &= (x_{i0}, v_{i0}).\end{aligned}\tag{12}$$

To reduce the number of equations in (12), we introduce the spatial and velocity differences:

$$x := x_1 - x_2, \quad v := v_1 - v_2.$$

Then, without loss of generality, we may assume

$$x_0 > 0, \quad v_0 > 0.\tag{13}$$

Note that the differences of x and v satisfy

$$\dot{x} = v, \quad \dot{v} = -K\psi(|x|)v,$$

or equivalently,

$$dv = -K\psi(|x|)dx.$$

We integrate the above relation to obtain

$$v(t) = v_0 - K \int_{x_0}^{x(t)} \psi(|y|)dy. \quad (14)$$

Depending on the relations between the coupling strength K and initial data, we might not have flocking in the sense of Definition 1. This negative result can be seen from the following proposition.

Proposition 1 [16] Suppose that the communication weight ψ takes the following form:

$$\psi(|x - y|) = \frac{1}{(1 + |x - y|)^\beta}, \quad \beta \geq 0, \quad (15)$$

and let (x, v) be the solution to the system (12)–(13) with initial data (x_0, v_0) . Then, the following assertions hold:

1. If (x_0, v_0) satisfies

$$v_0 = K \int_{x_0}^{\infty} \psi(|y|)dy, \quad (16)$$

then the positions of the two particles diverge with the same asymptotic velocities.

2. If (x_0, v_0) satisfies

$$v_0 > K \int_{x_0}^{\infty} \psi(|y|)dy, \quad (17)$$

then the positions of the two particles diverge with different asymptotic velocities.

Proof (i) Suppose (x_0, v_0) satisfies

$$v_0 = K \int_{x_0}^{\infty} \psi(|y|)dy.$$

Using (11), (14), and (16), we obtain

$$v(t) = v_0 - K \int_{x_0}^{x(t)} \psi(|y|)dy = K \int_{x(t)}^{\infty} \psi(|y|)dy > 0. \quad (18)$$

On the other hand, since $\frac{dx}{dt} = v > 0$ and $x(0) = x_0 > 0$, we have

$$x(t) > 0, \quad \text{i.e., } \psi(|x|) = \psi(x). \quad (19)$$

We now use (15), (18), and (19) to find a first-order equation for x :

$$\frac{dx}{dt} = K \int_{x(t)}^{\infty} \psi(y) dy = \frac{K}{\beta - 1} \frac{1}{(1 + x(t))^{\beta-1}}. \quad (20)$$

Directly integration (20) yields

$$x(t) = \left(\frac{\beta K t}{\beta - 1} + (1 + x_0)^{\beta} \right)^{1/\beta} - 1, \quad v(t) = \frac{K}{\beta - 1} \left(\frac{\beta K t}{\beta - 1} + (1 + x_0)^{\beta} \right)^{1/\beta-1}.$$

The above explicit formula implies

$$\lim_{t \rightarrow \infty} x(t) = \infty, \quad \lim_{t \rightarrow \infty} v(t) = 0.$$

Note that the velocity difference v goes to zero at the rate of $t^{-(1-1/\beta)}$.

(ii) Suppose (x_0, v_0) satisfies (17). It follows from (14) that

$$\begin{aligned} v(t) &= v_0 - K \int_{x_0}^{x(t)} \psi(|y|) dy \\ &= v_0 - K \int_{x_0}^{\infty} \psi(|y|) dy + K \int_{x(t)}^{\infty} \psi(|y|) dy. \end{aligned} \quad (21)$$

Note that (21) implies

$$v(t) \geq v_0 - K \int_{x_0}^{\infty} \psi(|y|) dy > 0, \quad t \geq 0.$$

Thus, the asymptotic velocities are not equal. On the other hand, if we set

$$v_{\infty} := v_0 - K \int_{x_0}^{\infty} \psi(|y|) dy,$$

then (21) implies

$$\frac{dx}{dt} = v_{\infty} + \frac{K}{\beta - 1} (1 + x(t))^{1-\beta}.$$

Clearly, $x(t)$ increases faster than $v_{\infty} t$ by the comparison theorem. This completes the proof.

Remark 2 1. It follows from Proposition 1 that even for a simple two-body system, the flocking theorem is not always true and it depends on the interplay between the

coupling strength and initial data. In the following three subsections, we present three variants of the C-S model with metric-dependent communications and discuss its flocking estimates.

2. The coupling strength function (15) appears also in different forms in the literature:

$$\psi(|x - y|) : \frac{1}{(1 + |x_i(t) - x_j(t)|^2)^{\frac{\beta}{2}}} \quad \text{or} \quad \frac{1}{(1 + |x_i(t) - x_j(t)|)^{\beta}}.$$

They are in fact equivalent to each other. Thus, we may use $\beta \leftrightarrow \frac{\beta}{2}$ in the models and results interchangeably.

3.1 General Symmetric Weights

In this subsection, we briefly review sufficient conditions for the emergence of asymptotic flocking for the C-S model in (9)–(11). Flocking estimate was first studied by Cucker and Smale [31]. They provided a sufficient condition on the formation of flocking for an algebraically decaying communication weight $\psi(r) = (1 + r^2)^{-\beta/2}$ with $\beta \geq 0$. For the short-ranged communication weight, they showed that asymptotic flocking is possible for initial configurations close to the flocking state using the self-bounding argument. Later, Cucker and Smale's results were further generalized to general nonincreasing communication weights (11) using a simpler energy method and Lyapunov functional approach, which were based on the ℓ^2 -norm and mixed $\ell^\infty - \ell^2$ norms in [1, 41, 42]. For a given configuration $(x, v) \in \mathbb{R}^{2dN}$ with a zero sum condition,

$$\sum_{i=1}^N x_i(t) = 0, \quad \sum_{i=1}^N v_i(t) = 0, \quad t \geq 0,$$

we set

$$|x|_\infty := \max_{1 \leq i \leq N} |x_i|, \quad |v|_\infty := \max_{1 \leq i \leq N} |v_i|.$$

Then, the norms $|x|_\infty$ and $|v|_\infty$ are Lipschitz continuous and satisfy a system of dissipative differential inequalities:

$$\left| \frac{d}{dt} |x|_\infty \right| \leq |v|_\infty, \quad \frac{d}{dt} |v|_\infty \leq -K\psi(2|x|_\infty) |v|_\infty \quad \text{a.e. } t \in (0, \infty). \quad (22)$$

Note that once we have a uniform upper bound for $|x|_\infty$, the second relation in (22) yields the exponential decay of $|v|_\infty$. Thus, we introduce Lyapunov-type functionals $\mathcal{L}_\pm(t) \equiv \mathcal{L}_\pm(x(t), v(t))$:

$$\mathcal{L}_\pm(t) := |v(t)|_\infty \pm \frac{K}{2} \int_0^{2|x(t)|_\infty} \psi(s) ds, \quad t \geq 0. \quad (23)$$

Then, it is easy to see the nonincreasing property of \mathcal{L}_\pm using (22):

$$\mathcal{L}_\pm(t) \leq \mathcal{L}_\pm(0), \quad t \geq 0,$$

which leads to the stability estimate of $\mathcal{L}_\pm(t)$:

$$|v(t)|_\infty + \frac{K}{2} \left| \int_{2|x_0|_\infty}^{2|x(t)|_\infty} \psi(s) ds \right| \leq |v_0|_\infty, \quad t \geq 0.$$

The following theorem is most relevant result on the flocking estimate.

Theorem 1 [1, 31, 41, 42] *Let (x, v) be a solution to (9)–(11) with initial data (x_0, v_0) satisfying the following condition:*

$$|x_0|_\infty > 0, \quad |v_0|_\infty < \frac{K}{2} \int_{|x_0|_\infty}^\infty \psi(2r) dr. \quad (24)$$

Then, there exists a positive number x_M such that

$$\sup_{t \geq 0} |x(t)| \leq x_M, \quad |v(t)| \leq |v_0| e^{-\psi(2x_M)t}, \quad t \geq 0.$$

Remark 3 1. The result of Theorem 1 can be restated as follows. For a given initial data (x_0, v_0) , there exists a coupling strength $K^*(x_0, v_0) =: 2|v_0|_\infty / \int_{|x_0|_\infty}^\infty \psi(2r) dr$ such that if $K > K^*(x_0, v_0)$, then we have an exponential flocking. Thus, natural question is to see the large-time dynamics in the regime $K < K^*(x_0, v_0)$ which Theorem 1 cannot be applied for. In this small coupling strength regime, there might be local flocking(or multicluster flocking). Recently, this issue has been addressed in a series of papers [16, 17, 39, 40].

2. In [1], C-S model with a singular communication weight is considered. Under certain condition of the initial configurations, the collision avoidance between agents is provided. Later, these conditions are refined in [12].

3.2 Nonsymmetric Interactions

After Cucker–Smale’s seminal work [32], one interesting extension of the C-S model has been proposed by Motsch and Tadmor in [52]. They replaced the symmetric interaction potential ψ_{ij} in the C-S model by a nonsymmetric one:

$$\psi_{ij} \longrightarrow \frac{\psi_{ij}}{\sum_{k=1}^N \psi_{ik}}.$$

Thus, the general C-S model proposed by Motsch and Tadmor reads as follows:

$$\begin{aligned}\frac{dx_i}{dt} &= v_i, \quad t > 0, \quad i = 1, \dots, N, \\ \frac{dv_i}{dt} &= \frac{K}{\sum_{k=1}^N \psi(|x_k - x_i|)} \sum_{j=1}^N \psi(|x_j - x_i|)(v_j - v_i).\end{aligned}\tag{25}$$

This model does not only take into account the distance between agents, but instead, the influence between agents is scaled in terms of their relative distance. Hence, it does not involve any explicit dependence on the number of agents. However, this extension of communication weight destroys the symmetry property of the original C-S model. The symmetry property of the communication weights is essential in energy estimate for the C-S model. Fortunately, the Lyapunov-type functional approach introduced in the previous subsection works for this nonsymmetric situation. To state their result, we introduce diameters for x and v :

$$D(x) := \max_{1 \leq i, j \leq N} |x_i - x_j|, \quad D(v) := \max_{1 \leq i, j \leq N} |v_i - v_j|.$$

Then, by the detailed calculations, they showed that these diameters satisfy a system of dissipative differential inequalities:

$$\left| \frac{dD(x)}{dt} \right| \leq D(v), \quad \frac{dD(v)}{dt} \leq -K\psi^2(D(x))D(v), \quad \text{a.e., } t \in (0, \infty).$$

Then, by using the idea of Lyapunov functional approach (23) depicted in the previous subsection, they obtain the following flocking estimate.

Theorem 2 [52] Suppose that ψ is positive and initial data satisfy

$$D(v_0) < \int_{D(x_0)}^{\infty} \psi^2(r)dr.$$

Then, the model (25) exhibits an asymptotic flocking:

$$\sup_{0 \leq t < \infty} D(x(t)) < \infty \quad \text{and} \quad \lim_{t \rightarrow \infty} D(v(t)) = 0.$$

In particular, if ψ has a fat tail such that

$$\int_c^{\infty} \psi^2(r)dr = \infty, \quad \text{for any positive } c,$$

then asymptotic flocking occurs for any initial data.

3.3 Bonding Force

For the realistic applications to robotic multiagent systems, we need to consider the formation control and collision avoidance. For this, we extend the C-S model by introducing additional interaction terms between agents in order to incorporate collision avoidance between agents and at the same time achieve tighter spatial configurations. We make use of not only position but also velocity information of the agents in order to derive the additional interaction between agents. This results in a control term which drives agents together or away in such a manner that the distance between agents converge to a nonzero constant value. The C-S model with aforementioned formation control and collision avoidance terms reads as follows:

$$\begin{aligned}\dot{x}_i &= v_i, \quad t > 0, \quad i = 1, \dots, N, \\ \dot{v}_i &= \frac{K_0}{N} \sum_{j=1}^N \psi(|x_j - x_i|)(v_j - v_i) + \frac{K_1}{N} \sum_{j=1}^N \frac{\langle v_j - v_i, x_j - x_i \rangle}{|x_j - x_i|} (x_j - x_i) \\ &\quad + \frac{K_2}{N} \sum_{j=1}^N (|x_j - x_i| - 2R)(x_j - x_i),\end{aligned}\tag{26}$$

where K_0 , K_1 , and K_2 are nonnegative coupling constants. Due to the translation invariance of (26), without loss of generality, we assume that

$$\sum_{i=1}^N x_i(t) = 0, \quad \sum_{i=1}^N v_i(t) = 0, \quad t \geq 0.\tag{27}$$

We define energy functionals:

$$\mathcal{E} := \mathcal{E}_k + \mathcal{E}_p, \quad \mathcal{E}_k := \frac{1}{2} \sum_{i=1}^N |v_i|^2, \quad \mathcal{E}_p := \frac{K_2}{4N} \sum_{1 \leq i, j \leq N} (|x_j - x_i| - 2R)^2,$$

where \mathcal{E}_k and \mathcal{E}_p represent kinetic and potential energies, respectively. Then, it follows from the energy estimates that the total energy \mathcal{E} satisfies dissipation estimate.

Proposition 2 [56] *For some $T \in (0, \infty]$, let (x, v) be a solution to (26)–(27) in the time interval $[0, T]$. Then, the energy functional \mathcal{E} is nonincreasing in time t :*

$$\mathcal{E}(t) + \int_0^t \mathcal{P}(\tau) d\tau = \mathcal{E}(0), \quad t \geq 0,$$

where energy production functional \mathcal{P} is given by the relation:

$$\mathcal{P} := \frac{K_0}{2N} \sum_{1 \leq i, j \leq N} \psi(|x_j - x_i|) |v_j - v_i|^2 + \frac{K_1}{2N} \sum_{1 \leq i, j \leq N} \left(\frac{d}{dt} |x_j - x_i|^2 \right).$$

This yields the flocking estimate for (26).

Theorem 3 [1, 56] Suppose that the communication weight ψ and initial data satisfy the following conditions.

$$\psi(r) \geq 0, \quad \exists r_0 \in (0, \infty] \text{ such that } \psi(r) > 0, \text{ for } r \leq r_0,$$

$$\psi_m := \min \left\{ \psi(r) : 0 \leq r \leq 2R + \sqrt{\frac{2N\mathcal{E}(0)}{K_2}} \right\} > 0, \quad \mathcal{E}(0) < K_2 R^2 N.$$

Let (x, v) be a global solution to (26)–(27). Then, the following assertions hold.

1. Asymptotic flocking occurs.

$$\sup_{0 \leq t < \infty} |x_i(t) - x_j(t)| < 2R + \sqrt{\frac{2N\mathcal{E}(0)}{K_2}}, \quad \lim_{t \rightarrow \infty} |v_i(t) - v_j(t)| = 0, \quad 1 \leq i, j \leq N.$$

2. The collision avoidance is guaranteed.

$$\inf_{0 \leq t < \infty} |x_i(t) - x_j(t)| > 0, \quad 1 \leq i, j \leq N.$$

4 Discrete-time Cucker–Smale Models with Leadership

In this section, we present the discrete-time C-S model. Let (x_i, v_i) denote the position and velocity of the i th particle, respectively, then the discrete-time C-S model, with time step $h > 0$, is governed by

$$\begin{aligned} x_i(t+1) &= x_i(t) + h v_i(t), \quad i = 1, 2, \dots, N, \\ v_i(t+1) &= v_i(t) + h \sum_{j=1}^N \phi(|x_i - x_j|) (v_j(t) - v_i(t)), \\ \phi(|x_i - x_j|) &= \frac{1}{(1 + |x_i(t) - x_j(t)|^2)^\beta}, \quad \beta \geq 0. \end{aligned} \tag{28}$$

The choice of weight function is a crucial ingredient which makes the C-S model attractive: The convergence results depend on conditions on the initial state only. In contrast, the convergence results in [44] for the linearized Vicsek model rely on some assumptions on the infinite time sequence of states. In the original C-S model, the interactions are bidirectional and thus symmetric. With bidirectional couplings, they used the Fiedler number of the symmetric Laplacian matrix to develop some

estimates on the iterates of the fluctuations of position and velocity so that the self-bounding lemma [31, Lemma 2] can be applied. The pioneering work [31] of Cucker and Smale gave the following flocking theorem.

Theorem 4 [31] *Consider the model (28) (which is under all-to-all and symmetric coupling). If $\beta < \frac{1}{2}$, the flocking occurs for any initial data; if $\beta \geq \frac{1}{2}$, the flocking occurs depending on the initial data.*

An example in [31] shows that when $\beta > \frac{1}{2}$, the unconditional flocking is not true. Thus, the exponent $\frac{1}{2}$ is regarded as the critical exponent for the unconditional flocking. In this section, we will briefly introduce some results on discrete-time C-S model with interactions under some leadership. Before we discuss the variant interaction topologies, we first briefly introduce some concepts in graph theory [34]. A digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (without self-loops) representing $(N + 1)$ particles with interaction in C-S model is defined by

$$\mathcal{V} := \{0, 1, \dots, N\}, \quad \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} \setminus \{(i, i) : i \in \mathcal{V}\}.$$

We say $(j, i) \in \mathcal{E}$ if and only if j influences i . In this case, we also write $j \in \mathcal{L}(i)$. The graph \mathcal{G} can be regarded as the information flowchart of a network structure; that is, we write

$$j \rightarrow i \iff (j, i) \in \mathcal{E}.$$

A directed path from j to i (of length $n + 1$) comprises a sequence of distinct arcs of the form $j \rightarrow k_1 \rightarrow k_2 \rightarrow \dots \rightarrow k_n \rightarrow i$. The distance from j to i is the length of the shortest path from j to i .

4.1 Hierarchical Leadership

The general form of a discrete-time C-S model is given by

$$\begin{aligned} x_i(t+1) &= x_i(t) + h v_i(t), \quad i = 0, 1, \dots, N, \\ v_i(t+1) &= v_i(t) + h \sum_{j=0}^N \phi_{ij}(x(t)) (v_j(t) - v_i(t)), \\ \phi_{ij}(x(t)) &= \begin{cases} 0 & \text{if } j \notin \mathcal{L}(i), \\ \phi(|x_i - x_j|) & \text{if } j \in \mathcal{L}(i). \end{cases} \end{aligned} \tag{29}$$

Here, $\mathcal{L}(i) \subset \{0, 1, \dots, N\}$, regarded as the *leader set* of agent i , is the set of agents which influence i directly. Thus, the interaction topology of the C-S model is registered in the configuration of $\mathcal{L}(i)$, or equivalently, the adjacency matrix $\Phi_x := (\phi_{ij}(x))$.

Shen extended the C-S flocking to an asymmetric structure in which the interactions are unidirectional. More precisely, he considered a C-S model under hierarchical

leadership, which means that the agents can be partially ordered in such a way that lower-rank agents are led and only led by some agents of higher ranks. The formal definition of hierarchical leadership is as follows.

Definition 2 [60] An $(N + 1)$ -flock $\{0, 1, \dots, N\}$ is said to be under hierarchical leadership if the following two statements hold:

- (a) $j \in \mathcal{L}(i)$ implies that $j < i$;
- (b) for any $i > 0$, $\mathcal{L}(i) \neq \emptyset$.

Definition 2 means that the adjacency matrix is triangular under a proper ordering of the agents. For the continuous-time model with hierarchy, Shen used the induction method to prove the unconditional flocking for $\beta < \frac{1}{2}$. The triangularity of the adjacency matrix is the key for the induction method; actually, the idea lies in the fact that the dynamics of agents $\{0, 1, \dots, N\}$ does not change if a new agent $N + 1$ ranking lowest is added. Later, Cucker and Dong extended the induction method to the discrete-time hierarchical model to improve the critical exponent (see [27]). The main result is as follows.

Theorem 5 [27, 60] Consider the model (29) with interaction topology as in Definition 2. If $\beta < \frac{1}{2}$, the flocking occurs for any initial data.

The induction method does not give any sufficient condition for the flocking behavior when β is no less than the critical exponent, i.e., $\beta \geq \frac{1}{2}$.

4.2 Individual Preference

A variant of the hierarchical C-S model is the flocking with individual preference [46]:

$$\begin{aligned} x_i(t+1) &= x_i(t) + hv_i(t), \quad i = 0, 1, \dots, N, \\ v_i(t+1) &= v_i(t) + h \sum_{j=0}^N \phi_{ij}(x(t)) (v_j(t) - v_i(t)) + h\delta_i(t)q_i(t), \\ \phi_{ij}(x(t)) &= \begin{cases} 0 & \text{if } j \notin \mathcal{L}(i), \\ H\phi(|x_i - x_j|), & \text{if } j \in \mathcal{L}(i). \end{cases} \end{aligned} \quad (30)$$

Here, the parameter $H > 0$ is incorporated as a measure of the strength of leader-follower interactions, $q_i(t) \in \mathbb{R}^3$ describes the temporarily preferred acceleration of agent i , and $\delta_i(t) \in \mathbb{R}$ is a local measure of the consensus at time t which determines the strength of preferred acceleration. As a special case, we may choose $q_i(t) \equiv \bar{q}_i$ for all time t ; then, it represents a constant preferred acceleration of agent i . In [46], a typical choice of δ_i depending on its relative velocities with respect to its leaders, which has been inspired by [28], was considered:

$$\delta_i(t) = \frac{1}{\#(\mathcal{L}(i))} \sum_{j \in \mathcal{L}(i)} |v_j(t) - v_i(t)|, \quad i = 1, 2, \dots, N, \quad \text{and} \quad \delta_0(t) \equiv 0.$$

Here, $\#(\mathcal{L}(i))$ denotes the cardinality of the leader set $\mathcal{L}(i)$. When an agent observes a consensus in its leaders and itself, it tends to give up its own preferred acceleration to follow the social leader–follower forces; otherwise, it will take an acceleration which is a combination of the social forces and its own preference; the strength of its preference is higher if it finds less consensus. On the other hand, one may assume $|q_i| \leq v$ for some $v \geq 0$ and all $1 \leq i \leq N$. Under these assumptions, the ratio $\frac{H}{v}$ expresses a trade-off between the social forces and individual preferences. Obviously, the terms $\delta_i(t)q_i(t)$ are state dependent even when $q_i(t)$'s are constant.

A simple example shows that the asymptotic flocking can fail due to a state-dependent individual preference (see [46, Example 2.1]). Thus, a natural question is whether it is possible to find an asymptotic flocking in the presence of such perturbations. The induction method works quite well for the hierarchical C-S flocking [27, 33, 60], but it cannot deal with the case of state-dependent perturbations.

In order to study such a system, one may consider the “fluctuation” system. Let

$$\begin{aligned} X &= (X_1, X_2, \dots, X_N)^\top := (x_1 - x_0, x_2 - x_0, \dots, x_N - x_0)^\top, \\ V &= (V_1, V_2, \dots, V_N)^\top := (v_1 - v_0, v_2 - v_0, \dots, v_N - v_0)^\top, \end{aligned} \quad (31)$$

and denote $Q(t) = (\delta_1(t)q_1(t), \delta_2(t)q_2(t), \dots, \delta_N(t)q_N(t))^\top$. Then, we use the C-S model (30) and hierarchical leadership to derive that

$$\begin{aligned} X(t+1) &= X(t) + hV(t), \\ V(t+1) &= P_t V(t) + hQ(t), \end{aligned}$$

where $P_t := I - hL_t$, regarded as the flocking matrix, is given by

$$P_t = \begin{pmatrix} 1 - hd_1(t) & 0 & \cdots & 0 \\ h\phi_{21}(t) & 1 - hd_2(t) & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ h\phi_{N1}(t) & h\phi_{N2}(t) & \cdots & 1 - hd_N(t) \end{pmatrix}.$$

Here, the matrices L_t and P_t are acting on $V(t) \in (\mathbb{R}^3)^N$ via the three dimensions individually. The crucial idea to deal with the perturbation $hQ(t)$ is a special matrix norm. For $\varepsilon \in (0, 1)$, we set an $N \times N$ diagonal matrix

$$D = D_\varepsilon := \begin{pmatrix} \varepsilon^{\ell(1)} & 0 & \cdots & 0 \\ 0 & \varepsilon^{\ell(2)} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \varepsilon^{\ell(N)} \end{pmatrix},$$

where $\ell(i)$ is the directed distance from the leader 0 to i , i.e., the number of edges in a shortest directed path from 0 to i . For any matrix $A \in \mathbb{R}^{N \times N}$, we define

$$\|A\|_\varepsilon := \|DAD^{-1}\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the infinity norm of matrices. For more detail, we refer to [46]. The crucial advantage of this norm is, for sufficient small time step h ,

$$\|P_t\|_\varepsilon \leq 1 - (1 - \varepsilon)h\phi_m(t) < 1,$$

where $\phi_m(t) := \min_{(i,j), j \in \mathcal{L}(i)} \phi_{ij}(x(t)) > 0$ (see [46, Proposition 3.1]). This well-chosen norm enables us to apply a bootstrapping self-bounding lemma and find sufficient conditions to guarantee the asymptotic flocking. For more detail, we refer to [46].

4.3 Rooted Leadership

A more general framework with leadership was introduced in [50]: the “rooted leadership” which requires that there exists a global leader which is not influenced by any other agent but influences them all either directly or indirectly.

Definition 3 [50] An $(N + 1)$ -flock $\{0, 1, \dots, N\}$ is said to be under *rooted leadership*, if there exists a root agent, say 0, which does not have an incoming path from others, whereas each agent in $\{1, 2, \dots, N\}$ has a directed path from 0.

In this case, the interactions can be unidirectional or bidirectional. Obviously, the hierarchical leadership is a special case of rooted leadership. The adjacency matrix is in general neither symmetric nor triangular, and the induction method cannot be applied. Using the same variable changes as in (31), we find a compact form

$$\begin{aligned} X(t+1) &= X(t) + hV(t), \\ V(t+1) &= P_t V(t), \end{aligned} \tag{32}$$

where $P_t := I - hL_t$. In [50], the *(sp)* matrix [67, 68] was employed to study the flocking of (32). As a result of the rooted leadership, the transition matrix P_t turns into an *(sp)* matrices when the time step h is small. Based on this observation, the authors could use the infinity norm to obtain an estimate on the iteration of transition matrices. This idea, together with the nice self-bounding lemma, leads to the following result.

Theorem 6 [50] Consider the model (29) with interaction topology as in Definition 3. If $\beta < \frac{1}{2L}$, the flocking occurs for any initial data; if $\beta \geq \frac{1}{2L}$, the flocking occurs depending on the initial data.

Remark 4 The parameter L , referred as the “depth” of the graph, is the largest distance (in the sense of graph theory) from the leader to other agents. Note that the conditions are sufficient but not necessary; thus, the critical exponent for this case is still open.

The scenario of leadership can be observed in many physical systems, e.g., flocks of flying birds and moving herds and governmental or military leadership. However, the connectivity topology might change over time. For example, in the movement of birds flock, some individuals fly so far away from the others that they cannot see each other from time to time. In social networks, it is more realistic to assume that the neighboring agents keep in connection only for a sequence of time slices rather than at all time instants. In [48], such an extended framework with joint rooted leadership was considered.

Definition 4 [48] The system is under *joint rooted leadership* across the time interval $[t_1, t_2]$, $(t_1, t_2 \in \mathbb{N}, t_1 < t_2)$ if for the union graph of $\{\mathcal{G}_{\sigma(t_1)}, \mathcal{G}_{\sigma(t_1+1)}, \dots, \mathcal{G}_{\sigma(t_2-1)}\}$, the agent 0 does not have an incoming path from others, whereas each agent in $\{1, 2, \dots, N\}$ has a directed path from 0.

Concerning the potential applications in engineering, this is relevant in at least two aspects. First, the failure of connections is very common due to the obstacles, faults, disturbances, and noise, and the joint connectivity helps the system to endure such failures which can be recovered after a finite recovery time. Second, it is relevant to the communication costs because more connections entail higher costs. In [48], a flocking result was established for joint rooted leadership, which says that the unconditional flocking occurs for $\beta < \frac{1}{2NT_0}$ where T_0 is the maximum length of the time intervals for the joint connectivity.

4.4 Alternating Leaders

In the previous studies, the leader agent is assumed to be fixed in temporal evolution of flocks. This is not realistic in some situations. For example, the dynamic leader–follower relation in pigeon flocks was discussed in [53]. Actually, we can often observe that the leaders can be changed during the migration of a migrating flock of birds. Of course, we can also find alternating leaders in our human social systems, for example, the periodic election of political leaders. In [47], the flocking with alternating leaders was discussed. We use $\{1, 2, \dots, m\}$ to label the admissible neighbor graphs with rooted leadership, and then, we write the system with a switching signal $\sigma : \mathbb{N} \rightarrow \{1, 2, \dots, m\}$ as follows:

$$\begin{aligned}
x_i(t+1) &= x_i(t) + hv_i(t), \quad i = 1, 2, \dots, N, \\
v_i(t+1) &= v_i(t) + h \sum_{j=1}^N \chi_{ij}^{\sigma(t)} \phi(|x_i - x_j|) [v_j(t) - v_i(t)], \\
\chi_{ij}^{\sigma(t)} &= \begin{cases} 0, & \text{if } j \notin \mathcal{L}^{\sigma(t)}(i), \\ 1, & \text{if } j \in \mathcal{L}^{\sigma(t)}(i). \end{cases}
\end{aligned} \tag{33}$$

Definition 5 [47] The system is under *rooted leadership with alternating leaders*, if the system is under rooted leadership at each time slice, but the leader agent, denoted by r_t , is dependent on time t .

Note that for the flocking with symmetric interactions or a fixed leader, such as [27, 31, 32, 41, 46, 48, 50, 60], the asymptotic velocity is a priori known, either the average of the initial velocities or just that of the leader. Thus, we can consider the dynamics of the fluctuations around the average velocity, or around the fixed leader, to study the flocking behavior. However, in the case of alternating leaders, one cannot a priori know the asymptotic velocity. To overcome this difficulty, one may combine the original system and a reference system. Let

$$\begin{aligned}
\hat{x} &:= (\hat{x}_1, \dots, \hat{x}_{N-1})^\top = (x_1 - x_N, \dots, x_{N-1} - x_N)^\top, \\
\hat{v} &:= (\hat{v}_1, \dots, \hat{v}_{N-1})^\top = (v_1 - v_N, \dots, v_{N-1} - v_N)^\top,
\end{aligned}$$

which satisfy, by (33), the reference system

$$\hat{x}(t+1) = \hat{x}(t) + h\hat{v}(t), \quad \hat{v}(t+1) = P_{\sigma(t)}\hat{v}(t). \tag{34}$$

On the other hand, a compact form of (33) reads

$$\begin{aligned}
x(t+1) &= x(t) + hv(t), \\
v(t+1) &= (I - hL_{\sigma(t)})v(t) =: F_{\sigma(t)}v(t).
\end{aligned} \tag{35}$$

In [29, 30], the convergence estimate for the first-order consensus model was studied. The self-bounding argument enables us to use their estimates in [29, 30] to find a priori estimate for $v(t)$ in (35). This easily turns into an estimate for $\hat{v}(t)$. Then, the self-bounding argument can be applied on system (34). The flocking result is as follows.

Theorem 7 [47] Consider the model (29) with interaction topology as in Definition 5. If $2\beta(N-1)^2 < 1$, the flocking occurs for any initial data; if $2\beta(N-1)^2 \geq 1$, the flocking occurs depending on the initial data.

Remark 5 In [29, 30], the authors studied the exponential consensus with more general interaction topologies, i.e., the rooted graph or joint rooted graph with a switch. Here, a *rooted graph* means it has at least one spanning tree. Thus, the rooted leadership turns into a special case of rooted graph and the case of alternating

leaders is a special case of rooted graphs undergoing a switch. We note that the methodology in [47] can be easily extended to the C-S model with such a general interaction topologies, i.e., the rooted graph or even the joint rooted graphs with a switch. Indeed, one can easily combine the consensus estimates in [29, 30] with the argument in [47] to cover the general interaction topologies, slightly changing the sufficient conditions.

5 Kinetic Description of Cucker–Smale Model

In this section, we first briefly present the derivation of the mean-field kinetic equation (2) from the particle system (9) using the BBGKY hierarchy in statistical mechanics. We also discuss interactions between flocking particles and fluid.

5.1 Derivation of the Kinetic C-S Model

Let us denote $f^N = f^N(x_1, \xi_1, \dots, x_N, \xi_N, t)$ by the N -particle probability density function. Note that the density function f^N is symmetric in its phase space arguments, i.e.,

$$f^N(\dots, x_i, \xi_i, \dots, x_j, \xi_j, \dots, t) = f^N(\dots, x_j, \xi_j, \dots, x_i, \xi_i, \dots, t),$$

due to the indistinguishability of particles.

We deduce from the conservation of mass that the time evolution of f^N can be written in the following form of Liouville equation:

$$\partial_t f^N + \sum_{i=1}^N \xi_i \cdot \nabla_{x_i} f^N + \frac{1}{N} \sum_{i=1}^N \nabla_{\xi_i} \cdot \left(\sum_{j=1}^N \psi(|x_i - x_j|) (\xi_j - \xi_i) f^N \right) = 0. \quad (36)$$

We next define the marginal distribution $f^N = f^N(x_1, \xi_1, t)$ as

$$f^N(x_1, \xi_1, t) := \int_{\mathbb{R}^{2d(N-1)}} f^N(x_1, \xi_1, x_-, \xi_-, t) dx_- d\xi_-,$$

where

$$(x_-, \xi_-) := (x_2, \xi_2, \dots, x_N, \xi_N).$$

Integrating the equation (36) with respect to $dx_- d\xi_-$, we find that the transport part and the forcing term of (36) can be estimated as

$$\int_{\mathbb{R}^{2d(N-1)}} \sum_{i=1}^N \xi_i \cdot \nabla_{x_i} f^N dx_- d\xi_- = \nabla_{x_1} \int_{\mathbb{R}^{2d(N-1)}} f^N dx_- d\xi_- = \xi_1 \cdot \nabla_{x_1} f^N (x_1, \xi_1, t)$$

and

$$\begin{aligned} & \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}^{2d(N-1)}} \sum_{j=1}^N \nabla_{\xi_i} \cdot (\psi(|x_i - x_j|) (\xi_j - \xi_i) f^N) dx_- d\xi_- \\ &= \frac{1}{N} \int_{\mathbb{R}^{2d(N-1)}} \sum_{2 \leq j \leq N} \nabla_{\xi_1} \cdot (\psi(|x_1 - x_j|) (\xi_j - \xi_1) f^N) dx_- d\xi_-, \end{aligned} \quad (37)$$

respectively. On the other hand, the symmetry property that for $j = 2, 3, \dots, N$

$$\begin{aligned} & \int_{\mathbb{R}^{2d(N-1)}} \psi(|x_1 - x_2|) (\xi_2 - \xi_1) f^N dx d\xi_- \\ &= \int_{\mathbb{R}^{2d(N-1)}} \psi(|x_1 - x_3|) (\xi_3 - \xi_1) f^N dx_- d\xi_- \end{aligned}$$

allows us to estimate (37) as follows.

$$\frac{1}{N} (N-1) \int_{\mathbb{R}^{2d(N-1)}} \psi(|x_1 - x_2|) \nabla_{\xi_1} \cdot ((\xi_2 - \xi_1) f^N) dx_- d\xi_- \quad (38)$$

We now define the two-particle marginal function g^N as

$$g^N(x_1, \xi_1, x_2, \xi_2, t) = \int_{\mathbb{R}^{2d(N-2)}} f^N dx_3 d\xi_3 \dots dx_N d\xi_N.$$

Then, by using the newly defined marginal function g^N , we rewrite (38) as

$$\left(1 - \frac{1}{N}\right) \nabla_{\xi_1} \cdot \int_{\mathbb{R}^{2d}} \psi(|x_1 - x_2|) (\xi_2 - \xi_1) g^N dx_2 d\xi_2.$$

Hence, we have

$$\partial_t f^N + \xi_1 \cdot \nabla_{x_1} f^N + \left(1 - \frac{1}{N}\right) \nabla_{\xi_1} \cdot \int_{\mathbb{R}^{2d}} \psi(|x_1 - x_2|) (\xi_2 - \xi_1) g^N dx_2 d\xi_2 = 0.$$

Then, by taking the mean-field limit $N \rightarrow \infty$ together with the following notations:

$$f(x_1, \xi_1, t) := \lim_{N \rightarrow \infty} f^N(x_1, \xi_1, t), \quad g(x_1, \xi_1, x_2, \xi_2, t) := \lim_{N \rightarrow \infty} g^N(x_1, \xi_1, x_2, \xi_2, t),$$

we obtain that the limiting functions f and g satisfy

$$\partial_t f + \xi_1 \cdot \nabla_{x_1} f + \nabla_{\xi_1} \cdot \int_{\mathbb{R}^{2d}} \psi(|x_1 - x_2|) (\xi_2 - \xi_1) g \, dx_2 \, d\xi_2 = 0.$$

In order to close the above equation, we use the following assumption called *molecular chaos*:

$$g(x_1, \xi_1, x_2, \xi_2, t) = f(x_1, \xi_1, t) f(x_2, \xi_2, t).$$

Finally, we relabel the position and velocity parameters, $(x_1, \xi_1) \mapsto (x, \xi)$ and $(x_2, \xi_2) \mapsto (y, \xi_*)$, and conclude the one-particle distribution function $f(x, \xi, t)$ satisfies the following Vlasov-type equation:

$$\begin{aligned} \partial_t f + \xi \cdot \nabla_x f + \nabla_\xi \cdot (F_a(f)f) &= 0, \quad (x, \xi) \in \mathbb{R}^d \times \mathbb{R}^d, \quad t > 0, \\ F_a(f)(x, \xi, t) &:= K \int_{\mathbb{R}^d \times \mathbb{R}^d} \psi(|x - y|) (\xi_* - \xi) f(y, \xi_*, t) dy d\xi_*. \end{aligned} \tag{39}$$

We can also adapt the classical result of [35] to rigorously derive the kinetic C-S equation (39) due to the smoothness of the communication weight ψ . More precisely, let us consider the empirical measure $\mu^N(t)$:

$$\mu^N(t) = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i(t), v_i(t))}, \tag{40}$$

where $(x_i(t), v_i(t))$ is a solution to the particle system (9). Then, we can show that μ^N satisfies the equation (39) in the sense of distributions, i.e., μ^N and f satisfy the same equation. Before stating the mean-field limit result, we introduce several notations: Let us denote by $\mathcal{M}(\mathbb{R}^{2d})$ the set of positive Radon measures and fix $T > 0$. $d_{BL}(\rho_1, \rho_2)$ stands for the bounded and Lipschitz distance between two measures $\rho_1, \rho_2 \in \mathcal{M}(\mathbb{R}^{2d})$, i.e.,

$$d_{BL}(\rho_1, \rho_2) := \sup_{h \in \mathcal{S}} \left| \int_{\mathbb{R}^d \times \mathbb{R}^d} h \, d\rho_1 - \int_{\mathbb{R}^d \times \mathbb{R}^d} h \, d\rho_2 \right|,$$

where \mathcal{S} is given by

$$\mathcal{S} := \left\{ h : \mathbb{R}^{2d} \rightarrow \mathbb{R} : \|h\|_{L^\infty} \leq 1 \text{ and } \text{Lip}(h) := \sup_{x \neq y} \frac{|h(x) - h(y)|}{|x - y|} \leq 1 \right\}.$$

Theorem 8 [8, 41] Given $f_0 \in \mathcal{M}(\mathbb{R}^{2d})$ compactly supported, take a sequence of μ_0^N of measures of the form:

$$\mu_0^N = \frac{1}{N} \sum_{i=1}^N \delta_{(x_i(0), v_i(0))}$$

such that

$$\lim_{N \rightarrow \infty} d_{BL}(\mu_0^N, f_0) = 0.$$

Consider $\mu^N(t)$ the empirical measure (40) with initial data $(x_i(0), v_i(0))$. Then, we have

$$\lim_{N \rightarrow \infty} d_{BL}(\mu^N(t), f(t)) = 0 \text{ for } t \geq 0,$$

where f is the unique measure solution to the equation (39) with initial data f_0 .

Remark 6 1. We can determine the measure solution f as the push forward of the initial density f_0 through the flow map generated by $(v, F_a(f))$, i.e., for any $h \in \mathcal{C}_c^1(\mathbb{R}^d \times \mathbb{R}^d)$ and $t, s \geq 0$

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} h(x, \xi) f(x, \xi, t) dx d\xi = \int_{\mathbb{R}^d \times \mathbb{R}^d} h(X(0; t, x, \xi), \Xi(0; t, x, \xi)) f_0(x, \xi) dx d\xi,$$

where (X, Ξ) satisfy

$$\begin{aligned} \frac{d}{dt} X(t; s, x, \xi) &= \Xi(t; s, x, \xi), \quad X(s; s, x, \xi) = x, \\ \frac{d}{dt} \Xi(t; s, x, \xi) &= F_a(f)(X(t; s, x, \xi), \Xi(t; s, x, \xi), t), \quad \Xi(s; s, x, \xi) = \xi. \end{aligned}$$

2. In [10, 43], the mean-field limits of C-S-type equations with topological interactions and sharp sensitivity regions are studied. In particular, the strategy used in [10] can be applied to other models including nonlocal repulsive–attractive forces locally averaged over sharp vision cones. C-S model with a singular communication weight is considered in [1, 9, 12], however, this is no still available literature on the rigorous derivation of that kinetic equation.
3. Large-time behavior of solutions for the equation (39) is provided in [15, 41, 42]. Kinetic C-S-type equations corresponding to (9) and (25) with noises are treated in [18, 36] showing the global existence of classical solutions near the global Maxwellian and its large-time behavior.

5.2 Interactions Between Flocking Particles and Fluids

In this part, we discuss the interactions between particles and its environment, i.e., fluids. Most available literature for collective behavior deals with only the dynamics of self-organized particles as a closed system, i.e., interactions with fluids and external

force fields are often ignored. However, as we can easily imagine, the dynamics of self-organized particles can be strongly influenced by neighboring fluids and force fields, for example, water, gas, and electromagnetic waves. Thus, incorporating these neglected effects in the modeling of the self-organized particles will be necessary. In [3], the dynamics of flocking particles governed by the C-S model interacting with viscous compressible fluids through a drag forcing term are taken into account in the spatial periodic domain \mathbb{T}^3 . More precisely, let $f = f(x, \xi, t)$ be the one-particle distribution function of the C-S flocking particles at $(x, \xi) \in \mathbb{T}^3 \times \mathbb{R}^3$ and $n = n(x, t)$, $v = v(x, t)$ be the local mass density and bulk velocity of the isentropic compressible fluid, respectively. Then, the situation we mentioned above is governed by

$$\begin{aligned} \partial_t f + \xi \cdot \nabla_x f + \nabla_\xi \cdot (F_a(f)f + F_d(v)f) &= 0, \quad (x, \xi) \in \mathbb{T}^3 \times \mathbb{R}^3, \quad t > 0, \\ \partial_t n + \nabla_x \cdot (nv) &= 0, \\ \partial_t(nv) + \nabla_x \cdot (nv \otimes v) + \nabla_x p(n) + Lv &= - \int_{\mathbb{R}^3} F_d(v)f d\xi, \end{aligned} \tag{41}$$

where the pressure p and the Lamé operator L are given by

$$\begin{aligned} p(n) &= n^\gamma \quad \text{with } \gamma > 1, \\ Lv &= -\mu \Delta_x v - (\mu + \lambda) \nabla_x (\nabla_x \cdot v) \quad \text{with } \mu > 0 \quad \text{and } \lambda + 2\mu > 0. \end{aligned}$$

Here, we assumed the coupling strength $K = 1$, and F_a and F_d represent the alignment and the drag forces in velocities, respectively:

$$\begin{aligned} F_a(f)(x, \xi, t) &= \int_{\mathbb{T}^3 \times \mathbb{R}^3} \psi(|x - y|)(\xi_* - \xi)f(y, \xi_*, t) dy d\xi_* \quad \text{with } \psi \geq 0, \\ F_d(x, \xi, t) &= v(x, t) - \xi. \end{aligned}$$

Recently, this kind of coupled kinetic–fluid system describing the interactions between particles and fluid has received increasing attention due to a number of their applications in the field of, for example, biotechnology and medicine and in the study of sedimentation phenomenon, compressibility of droplets of the spray, cooling tower plumes, diesel engines, etc. [7, 59, 61, 65]. We refer to [54, 66] for more physical backgrounds of the modeling issues in a kinetic–fluid system.

In the lemma below, we present the properties of conservation and energy estimates for the system (41). For detail of the proof, we refer to [3].

Lemma 1 *Let (f, n, v) be a classical solution to the system (41). Then, we have*

(i) *Conservation of the mass:*

$$\frac{d}{dt} \int_{\mathbb{T}^3 \times \mathbb{R}^3} f \, dx d\xi = \frac{d}{dt} \int_{\mathbb{T}^3} n \, dx = 0.$$

(ii) *Conservation of the total momentum:*

$$\frac{d}{dt} \left(\int_{\mathbb{T}^3 \times \mathbb{R}^3} \xi f \, dx d\xi + \int_{\mathbb{T}^3} nv \, dx \right) = 0.$$

(iii) *Dissipation of the total energy:*

$$\begin{aligned} & \frac{d}{dt} \frac{1}{2} \left(\int_{\mathbb{T}^3 \times \mathbb{R}^3} |\xi|^2 f \, dx d\xi + \int_{\mathbb{T}^3} n|v|^2 \, dx + \frac{2}{\gamma - 1} \int_{\mathbb{T}^3} n^\gamma \, dx \right) \\ &= - \int_{\mathbb{T}^6 \times \mathbb{R}^6} \psi(|x - y|) |\xi - \xi_*|^2 f(x, \xi) f(y, \xi_*) \, dy dx d\xi d\xi_* \\ & \quad - \int_{\mathbb{T}^3 \times \mathbb{R}^3} |v - \xi|^2 f \, dx d\xi. \end{aligned}$$

The global existence of unique strong solutions for the system (41) is studied in [3] under suitable assumptions on the initial data such as smallness and smoothness. For the large-time behavior of solutions to types of equations (41), in [19], the following Lyapunov functional \mathcal{L} measuring the fluctuation of momentum and mass from the averaged quantities is introduced:

$$\begin{aligned} \mathcal{L}(f, \rho, u) := & \int_{\mathbb{T}^3 \times \mathbb{R}^3} |\xi - \xi_c|^2 f \, dx d\xi + \int_{\mathbb{T}^3} n|v - j_c|^2 \, dx + \int_{\mathbb{T}^3} (n - n_c)^2 \, dx \\ & + |\xi_c - j_c|^2, \end{aligned}$$

where

$$\xi_c(t) := \frac{\int_{\mathbb{T}^3 \times \mathbb{R}^3} \xi f \, dx d\xi}{\int_{\mathbb{T}^3 \times \mathbb{R}^3} f \, dx d\xi}, \quad j_c(t) := \frac{\int_{\mathbb{T}^3} nv \, dx}{\int_{\mathbb{T}^3} n \, dx}, \quad f_c(t) := \int_{\mathbb{T}^3 \times \mathbb{R}^3} f \, dx d\xi,$$

and

$$n_c(t) := \int_{\mathbb{T}^3} n \, dx.$$

Theorem 9 [19] *Let (f, n, v) be a global classical solution to the system (41) satisfying*

- (i) $\|\rho_f\|_{L^\infty(\mathbb{R}_+; L^{3/2}(\mathbb{T}^3))} < \infty$ where $\rho_f(x, t) := \int_{\mathbb{R}^3} f(x, \xi, t) \, d\xi$,
- (ii) $n(x, t) \in [0, \bar{n}]$ for all $(x, t) \in \mathbb{T}^3 \times \mathbb{R}_+$ and $n_c(0) > 0$,
- (iii) $v \in L^\infty(\mathbb{T}^3 \times \mathbb{R}_+)$ and $E_0 > 0$ is small enough,

where E_0 is the initial total energy given by

$$E_0 := \int_{\mathbb{T}^3 \times \mathbb{R}^3} |\xi|^2 f_0 \, dx d\xi + \int_{\mathbb{T}^3} n_0 |v_0|^2 dx + \frac{2}{\gamma - 1} \int_{\mathbb{T}^3} n_0^\gamma \, dx.$$

Then, we have

$$\mathcal{L}(t) \leq C \mathcal{L}_0 e^{-\lambda t} \quad t \geq 0,$$

where C and λ are positive constants independent of t .

Theorem 9 shows the alignment between flocking particles and fluid velocities as time goes on exponentially fast. More precisely, it follows from conservations of masses and total momentum that

$$\begin{aligned} & \xi_c(t) - j_c(t) \\ &= (f_c(0) + 1)\xi_c(t) - \frac{1}{n_c(0)} \left(\int_{\mathbb{T}^3 \times \mathbb{R}^3} \xi f_0(x, \xi) \, dx d\xi - \int_{\mathbb{T}^3} n_0(x) v_0(x) \, dx \right). \end{aligned}$$

This yields

$$\xi_c(t), j_c(t) \rightarrow \frac{1}{n_c(0)(f_c(0) + 1)} \left(\int_{\mathbb{T}^3 \times \mathbb{R}^3} \xi f_0(x, \xi) \, dx d\xi - \int_{\mathbb{T}^3} n_0(x) v_0(x) \, dx \right),$$

as $t \rightarrow \infty$. We notice that it is natural to expect from the presence of the drag forcing term in the kinetic and fluid equations (41).

For the case when the fluid is incompressible, the global well-posedness and *a priori* estimate of large-time behaviors of solutions are studied in [2, 4, 5, 23, 24]. In particular, the density-dependent drag forcing term which is more physically relevant is considered in [25] and the global existence of strong solutions and large-time behavior is obtained. In [11], the dynamics of particles immersed in an incompressible fluid through local alignments are taken into account. Unlike the C-S alignment force F_a , each particle actively tries to align its velocity to that of its closest neighbors. For this system, the global existence of weak solutions, hydrodynamic limit corresponding to strong noise and local alignment, and large-time behavior of solutions is established. Very recently, the finite-time blowup phenomena of classical solutions to (41) and other related systems under suitable assumptions on the initial configurations are provided in [22].

6 Hydrodynamic Descriptions for Flocking Behavior

In this section, we discuss hydrodynamic models describing the flocking behavior of the C-S ensemble. We first deal with the hydrodynamic C-S model introduced in Section 2 and then discuss its coupling with isentropic Navier–Stokes equations via the drag force.

6.1 A Hydrodynamic Cucker–Smale Model

In this part, we discuss a hydrodynamic C-S model:

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho u) &= 0, \quad x \in \Omega, \quad t > 0 \\ \partial_t (\rho u) + \nabla \cdot (\rho u \otimes u) &= \int_{\Omega} \psi(|x - y|)(u(y) - u(x))\rho(x)\rho(y) dy, \end{aligned} \tag{42}$$

subject to initial density and velocity

$$(\rho(x, t), u(x, t))|_{t=0} = (\rho_0(x), u_0(x)) \quad x \in \Omega.$$

Here, we again assumed the coupling strength $K = 1$ for simplicity. Without loss of generality, we may assume that ρ is a probability density function, i.e., $\|\rho(\cdot, t)\|_{L^1} = 1$ since the total mass is conserved in time.

For the system (42), the global existence of classical solutions in periodic domain and moving boundary problem is studied in [37, 38] under suitable assumptions on the initial data and the communication weight. In one dimension, a complete description of the critical threshold to the system (42) leading to a sharp dichotomy condition between global-in-time existence or finite-time blowup of strong solutions is obtained in [13] which extends both the sub- and supercritical regions derived in [62]. Other interaction forces, such as attractive/repulsive forces in position, are also considered in [13] for the classification of the critical thresholds to the system (42).

Inspired by [15] (see also Section 3.1), we show the large-time behavior of solutions in L^∞ framework. For this, we first set spatial diameter R^x and velocity diameter R^u as follows:

$$R^x(t) := \sup_{x, y \in \text{supp } \rho(\cdot, t)} |x - y| \quad \text{and} \quad R^u(t) := \sup_{x, y \in \text{supp } \rho(\cdot, t)} |u(x, t) - u(y, t)|.$$

Using the above notations, we define the notion of flocking behavior for the system (42).

Definition 6 Let (ρ, u) be the solution to (42). Then, the system (42) exhibits global flocking if and only if the following two conditions hold.

- (i) The spatial diameter R^x is uniformly bounded in time, i.e., there exists a positive constant C which is independent of t such that

$$\sup_{t \geq 0} R^x(t) \leq C.$$

- (ii) The velocity diameter R^u decays to zero as time goes to infinity:

$$\lim_{t \rightarrow \infty} R^u(t) = 0.$$

Theorem 10 Let (ρ, u) be any smooth solutions to the system (42) with compactly supported initial data (ρ_0, u_0) . Suppose that the initial spatial and velocity diameters satisfy

$$R_0^u < \int_{R_0^x}^{\infty} \psi(s) ds.$$

Then, the system (42) exhibits the flocking behavior.

Proof Let us consider the following two characteristic flows:

$$\frac{dX(t)}{dt} = u(X(t), t) \quad \text{and} \quad \frac{dY(t)}{dt} = u(Y(t), t),$$

with the initial conditions $X(0) = x$ and $Y(0) = y$ where $x, y \in \text{supp } \rho_0$. For notational simplicity, in the rest of estimates, we omit the time dependence of X and u , i.e., $X := X(t)$ and $u(X) := u(X(t), t)$; similarly, it is also taken for the Y and $u(Y)$. Note that

$$\frac{du(X)}{dt} = (\partial_t + u \cdot \nabla_x)u = \int_{\mathbb{R}^d} \psi(|X - y|)(u(y) - u(X))\rho(y) dy \quad \text{on } \text{supp } \rho(t).$$

For the proof, it is enough to show that the spatial and velocity diameters satisfy the following differential inequalities:

$$\begin{aligned} \frac{d}{dt}R^x(t) &\leq R^u(t), \\ \frac{d}{dt}R^u(t) &\leq -\psi(R^x(t))R^u(t), \end{aligned} \tag{43}$$

due to Theorem 1. First, it easily follows from the definition of the R^x and R^v that

$$\frac{1}{2} \frac{d}{dt}|X - Y|^2 = (X - Y) \cdot (u(X) - u(Y)) \leq R^x R^u,$$

and this yields

$$\frac{d}{dt}R^x(t) \leq R^u(t).$$

Since we are dealing with the classical solutions, we can choose X and Y such that $R^u = |u(X) - u(Y)|$ and R^u is differentiable with respect to time almost everywhere. For the estimate of time evolution of R^u , we obtain

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} (R^u)^2 &= \frac{1}{2} \frac{d}{dt} |u(X) - u(Y)|^2 = (u(X) - u(Y)) \cdot (F(\rho)(X) - F(\rho)(Y)) \\ &=: J_1 + J_2, \end{aligned}$$

where

$$F(\rho)(X) := \int_{\mathbb{R}^d} \psi(|X - y|)(u(y) - u(X))\rho(y) dy.$$

For the estimate of J_1 , we use the fact

$$(u(X) - u(Y)) \cdot (u(z) - u(X)) = (u(X) - u(Y)) \cdot (u(z) - u(Y) + u(Y) - u(X)) \leq 0,$$

for $X, Y, z \in \text{supp } \rho(t)$, due to the choice of X and Y . This yields

$$\begin{aligned} J_1 &= \int_{\mathbb{R}^d} \psi(|X - z|) (u(X) - u(Y)) \cdot (u(z) - u(X)) \rho(z) dz \\ &\leq \psi(R^x) \int_{\mathbb{R}^d} (u(X) - u(Y)) \cdot (u(z) - u(X)) \rho(z) dz \end{aligned}$$

Similarly, we can find

$$J_2 \leq -\psi(R^x) \int_{\mathbb{R}^d} (u(X) - u(Y)) \cdot (u(z) - u(Y)) \rho(z) dz.$$

Hence, we have

$$\frac{1}{2} \frac{d}{dt} (R^u)^2 \leq -\psi(R^x) |u(X) - u(Y)|^2 = -\psi(R^x) (R^u)^2,$$

where we used

$$\int_{\mathbb{R}^d} \rho dx = 1.$$

This completes the proof.

6.2 Hydrodynamic Model for the Interaction of Cucker–Smale Flocking Particles and Fluids

By using a similar derivation presented in Section 6.1, we can also derive the two-phase fluid model consisting of the pressureless Euler equations and the isentropic Navier–Stokes equations where the coupling is through the drag force from the coupled kinetic–fluid system (41). More precisely, this hydrodynamic system is governed by

$$\begin{aligned} \partial_t \rho + \nabla_x \cdot (\rho u) &= 0, \quad x \in \mathbb{T}^3, \quad t > 0, \\ \partial_t(\rho u) + \nabla_x \cdot (\rho u \otimes u) &= -\rho(u - v) - \rho \int_{\mathbb{T}^3} \psi(|x - y|)(u(x) - u(y))\rho(y) dy, \\ \partial_t n + \nabla_x \cdot (nv) &= 0, \\ \partial_t(nv) + \nabla_x \cdot (nv \otimes v) + \nabla_x p(n) + Lv &= \rho(u - v). \end{aligned} \quad (44)$$

Here, $\rho(x, t)$ and $n(x, t)$ represent the particle density and the fluid density at a domain $(x, t) \in \mathbb{T}^3 \times \mathbb{R}_+$, and $u(x, t)$ and $v(x, t)$ represent the corresponding bulk velocities for $\rho(x, t)$ and $n(x, t)$, respectively.

For the global-in-time existence of classical solutions to the system (44), one of the main difficulties in analyzing it arises from the formation of singularities. We notice that the system (44) without the drag and nonlocal velocity alignment forces reduces to the pressureless Euler equations, and it is well known that the Euler equations may develop a singularity in finite time no matter how smooth the initial data are. For this reason, it is natural to extend the notion of solutions to the measure-valued solutions. Concerning this issue, an interesting question is whether the interactions with viscous fluids through the drag force can prevent the formation of the finite-time singularities and whether the system can admit the global classical solutions.

In [37], the global existence of classical solutions for the pressureless Euler/incompressible Navier–Stokes equations with the nonlocal alignment forces and its large-time behavior is studied. It is interesting that the (*a priori*) estimate of time behavior of solutions plays an important role in constructing the global-in-time solutions. For the system (44) without the alignment force, i.e., $\psi \equiv 0$, the global existence and uniqueness of classical solutions and *a priori* estimate of large-time behavior of solutions showing that the two fluid velocities are aligned exponentially fast are obtained in [26]. The strategy used in [26] can be directly applied to the system (44), and in particular, we can deduce from [26] the following *a priori* estimate for the large-time behavior of solutions to the system (44).

Theorem 11 *Let (ρ, u, n, v) be the classical solutions to the system (44) satisfying*

- $$\begin{aligned} (i) \quad &\rho, n, v \in L^\infty(\mathbb{T}^3 \times \mathbb{R}_+). \\ (ii) \quad &\rho_c(0), n_c(0) \in (0, \infty) \text{ and } \tilde{E}_0 > 0 \text{ is small enough,} \end{aligned} \quad (45)$$

where \tilde{E}_0 is an initial total energy given by

$$\tilde{E}_0 := \int_{\mathbb{T}^3} \rho_0 |u_0|^2 dx + \int_{\mathbb{T}^3} n_0 |v_0|^2 dx + \frac{2}{\gamma - 1} \int_{\mathbb{T}^3} n_0^\gamma dx.$$

Then, we have

$$\tilde{\mathcal{L}}(t) \leq C \tilde{\mathcal{L}}_0 e^{-\lambda t}, \quad t \in [0, T],$$

for some constants C and $\lambda > 0$, where

$$\tilde{\mathcal{L}}(t) = \int_{\mathbb{T}^3} \rho |u - m_c|^2 dx + \int_{\mathbb{T}^3} n |v - j_c|^2 dx + |m_c - j_c|^2 + \int_{\mathbb{T}^3} (n - n_c)^2 dx,$$

and

$$m_c(t) := \frac{\int_{\mathbb{T}^3} \rho u dx}{\int_{\mathbb{T}^3} \rho dx}.$$

It is worth noticing that we do not require that the $L^\infty(\mathbb{T}^3)$ -norms of solutions ρ , n , and v should be small, and we need only the small initial total energy.

As mentioned in Remark 1, we can derive the isothermal Euler equations coupled with Navier–Stokes equations from the kinetic–fluid system (41) by considering the strong noise and the local alignment instead of the velocity alignment force $F_a(f)$. We notice that for the hydrodynamic limit to be rigorously derived (and not only formally) within the framework of relative entropy techniques, one of the main challenges is to establish the global existence of strong solutions of the fluid equation. To be more precise, the standard argument for the hydrodynamic limit is based on the weak–strong stability employing a relative entropy functional and holds as long as there exist global weak solutions to the kinetic–fluid equations and strong solutions to the fluid–fluid equations. However, as we briefly mentioned as before, solutions of Euler-type equations are well known to possibly develop a singularity in a finite time no matter how smooth the initial data are.

For the isothermal Euler/incompressible Navier–Stokes equations, the global existence and uniqueness of classical solutions are studied in [20] by reinterpreting the drag forcing term as the relative damping and extracting the smoothing effect of viscosity in the Navier–Stokes equations. This yields that its rigorous derivation from Vlasov–Fokker–Planck/incompressible Navier–Stokes equations with local alignment forces for some particular regime of the dispersed phase obtained in [11] holds for all time. For the interactions with compressible fluids, i.e., isothermal Euler/compressible Navier–Stokes equations, the global-in-time existence of classical solutions and its large-time behavior are obtained in [21].

Acknowledgements The work of S.-Y. Ha was supported by the Samsung Science and Technology Foundation under Project Number SSTF-BA1401-03. The work of Y.-P. Choi was supported by Engineering and Physical Sciences Research Council (EP/K008404/1). The work of Z. Li was supported by the National Natural Science Foundation of China Grant No.11401135, and Fundamental Research Funds for the Central Universities (HIT.BRETIII.201501 and HIT.PIRS.201610).

References

1. Ahn, S., Choi, H., Ha, S.-Y., and Lee, H.: On the collision avoiding initial-congurations to the Cucker–Smale type flocking models, *Comm. Math. Sci.* **10**, 625–643 (2012).
2. Bae, H.-O., Choi, Y.-P., Ha, S.-Y., and Kang, M.-J.: Time-asymptotic interaction of flocking particles and incompressible viscous fluid, *Nonlinearity* **25**, 1155–1177 (2012).

3. Bae, H.-O., Choi, Y.-P., Ha, S.-Y., and Kang, M.-J.: Asymptotic flocking dynamics of Cucker-Smale particles immersed in compressible fluids, *Disc. and Cont. Dyn. Sys.* **34**, 4419–4458 (2014).
4. Bae, H.-O., Choi, Y.-P., Ha, S.-Y., and Kang, M.-J.: Global existence of strong solution for the Cucker-Smale-Navier-Stokes system, *J. Diff. Eqns.* **257**, 2225–2255 (2014).
5. Bae, H.-O., Choi, Y.-P., Ha, S.-Y., and Kang, M.-J.: Global existence of strong solutions to the Cucker-Smale-Stokes system, *J. Math. Fluid Mech.*, **18**, 381–396 (2016).
6. Ballerini, M., Cabibbo, N., Candelier, R., et al.: Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study, *Proc. Nat. Acad. Sci.* **105**, 1232–1237 (2008).
7. Boudin, L., Desvillettes, L., and Motte, R.: A modelling of compressible droplets in a fluid, *Comm. Math. Sci.* **1**, 657–669 (2003).
8. Cañizo, J. A., Carrillo, J. A., and Rosado, J.: A well-posedness theory in measures for some kinetic models of collective motion, *Math. Mod. Meth. Appl. Sci.* **21**, 515–539 (2011).
9. Carrillo, J. A., Choi, Y.-P., and Hauray, M.: Local well-posedness of the generalized Cucker-Smale model, *ESAIM: Proc.* **47**, 17–35 (2014).
10. Carrillo, J. A., Choi, Y.-P., Hauray, M., and Salem, S.: Mean-field limit for collective behavior models with sharp sensitivity regions, to appear in *J. Eur. Math. Soc.*
11. Carrillo, J. A., Choi, Y.-P., and Karper, T.: On the analysis of a coupled kinetic-fluid model with local alignment forces, *Annales de l'IHP-ANL*, **33**, (2016), 273–307.
12. Carrillo, J. A., Choi, Y.-P., and Peszek, J.: Sharp Conditions to avoid collisions in singular Cucker-Smale interactions, Preprint.
13. Carrillo, J. A., Choi, Y.-P., Tadmor, E., and Tan, C.: Critical thresholds in 1D Euler equations with nonlocal interaction forces, *Math. Mod. Meth. Appl. Sci.* **26**, 185–206 (2016).
14. Carrillo, J. A., Fornasier, M., Toscani, G. and Vecil, F.: Particle, kinetic and hydrodynamic models of swarming. In *Mathematical modeling of collective behavior in Socio-Economic and Life Sciences*, 297–336 (2010).
15. Carrillo, J. A., Fornasier, M., Rosado, J., and Toscani, G.: Asymptotic flocking dynamics for the kinetic Cucker-Smale model, *SIAM J. Math. Anal.* **42**, 218–236 (2010).
16. Cho, J., Ha, S.-Y., Huang, F., Jin, C. and D. Ko: Emergence of bi-cluster flocking for the Cucker-Smale model, *Math. Mod. Meth. Appl. Sci.* **14**, (2016) doi:[10.1142/S0219530515400023](https://doi.org/10.1142/S0219530515400023).
17. Cho, J., Ha, S.-Y., Huang, F., Jin, C. and D. Ko: Emergence of bi-cluster flocking for agent-based models with unit speed constraint. *Analysis and Applications* **14**, 39–73 (2016).
18. Choi, Y.-P.: Global classical solutions of the Vlasov-Fokker-Planck equation with local alignment forces, *Nonlinearity*, **29**, (2016), 1887–1916.
19. Choi, Y.-P.: Large-time behavior for the Vlasov/compressible Navier-Stokes equations, *J. Math. Phys.*, **57**, 071501, (2016).
20. Choi, Y.-P.: Compressible Euler equations interacting with incompressible flow, *Kinetic and Related Models* **8**, 335–358 (2015).
21. Choi, Y.-P.: Global classical solutions and large-time behavior of the two-phase fluid model, *SIAM J. Math. Anal.*, **48**, (2016), 3090–3122.
22. Choi, Y.-P.: Finite-time blow-up phenomena of Vlasov/Navier-Stokes equations and related systems, preprint.
23. Choi, Y.-P. and Lee, J.: Global existence of weak and strong solutions to Cucker-Smale-Navier-Stokes equations in \mathbb{R}^2 , *Nonlinear Anal.-Real.* **27**, 158–182 (2016).
24. Choi, Y.-P. and Kwon, B.: Two-species flocking particles immersed in a fluid, *Comm. Info. Sys.* **13**, 123–149 (2013).
25. Choi, Y.-P. and Kwon, B.: Global well-posedness and large-time behavior for the inhomogeneous Vlasov-Navier-Stokes equations, *Nonlinearity* **28**, 3309–3336 (2015).
26. Choi, Y.-P. and Kwon, B.: The Cauchy problem for the pressureless Euler/isentropic Navier-Stokes equations, *J. Diff. Eqns.*, **261**, 654–711 (2016).
27. Cucker, F., and Dong J.G.: On the critical exponent for flocks under hierarchical leadership, *Math. Mod. Meth. Appl. Sci.* **19**, 1391–1404 (2009).
28. Cucker, F., and Huepe, C.: Flocking with informed agents, *Maths. in Action* **1**, 1–25 (2008).

29. Cao, M., Morse, A. S., and Anderson, B. D. O.: Reaching a consensus in a dynamically changing environment: A graphic approach, *SIAM J. Control Optim.* **47**, 575–600 (2008).
30. Cao, M., Morse, A. S., and Anderson, B. D. O.: Reaching a consensus in a dynamically changing environment: Vonvergence rates, measurement delays, and asynchronous events, *SIAM J. Control Optim.* **47**, 601–623 (2008).
31. Cucker, F., and Smale S.: Emergent behavior in flocks, *IEEE Trans. Autom. Control* **52**, 852–862 (2007).
32. Cucker, F., and Smale S.: On the mathematics of emergence, *Japan. J. Math.* **2**, 197–227 (2007).
33. Dalmao, F., and Mordecki, E.: Cucker-Smale flocking under hierarchical leadership and random interactions, *SIAM J. Appl. Math.* **71**, 1307–1316 (2010).
34. Diestel, R.: Graph Theory, Graduate Texts in Mathematics New York, U.S.A.: Springer-Verlag, (1997).
35. Dobrushin, R.: Vlasov equations, *Funct. Anal. Appl.* **13**, 115–123, (1979).
36. Duan, R., Fornasier, M., and Toscani, G.: A kinetic flocking model with diffusion, *Commun. Math. Phys.* **300**, 95–145 (2010).
37. Ha, S.-Y., Kang, M.-J., and Kwon, B.: A hydrodynamic model for the interaction of Cucker-Smale particles and incompressible fluid, *Math. Mod. Meth. Appl. Sci.* **24**, 2311–2359 (2014).
38. Ha, S.-Y., Kang, M.-J., and Kwon, B.: Emergent dynamics for the hydrodynamic Cucker-Smale system in a moving domain, *SIAM. Math. Anal.* **47**, 3813–3831 (2015).
39. Ha, S.-Y., Ko, D., Zhang, Y. and Zhang, X.: Emergent dynamics in the interactions of Cucker-Smale ensembles, to appear in Kinetic and Related Models.
40. Ha, S.-Y., Ko, D. and Zhang, Y.: A criterion for non-flocking and emergence of multi-cluster flocking for the Cucker-Smale model, to appear in *Math. Mod. Meth. Appl. Sci.*
41. Ha, S.-Y. and Liu, J.-G.: A simple proof of Cucker-Smale flocking dynamics and mean field limit. *Comm. Math. Sci.* **7**, 297–325 (2009).
42. Ha, S.-Y. and Tadmor, E.: From particle to kinetic and hydrodynamic description of flocking, *Kinetic and Related Models* **1**, 415–435 (2008).
43. Haskovec, J.: Flocking dynamics and mean-field limit in the Cucker-Smale-type model with topological interactions, *Physica D* **261**, 42–51 (2013).
44. Jadbabaie, A., Lin, J., and Morse, A.: Coordination of groups of mobile autonomous agents using nearest neighbor rules, *IEEE Trans. Autom. Control* **48** 988–1001 (2003).
45. Karper, T. K., Mellet, A., and Trivisa, K.: Hydrodynamic limit of the kinetic Cucker-Smale flocking model **25**, 131–163 (2015).
46. Li, Z.: Effectual leadership in flocks with hierarchy and individual preference, *Disc. Cont. Dyn. Syst. A* **34**, 3683–3702 (2014).
47. Li, Z., and Ha, S.-Y.: On the Cucker-Smale flocking with alternating leaders, *Quart. Appl. Math.* **73**, 693–709 (2015).
48. Li, Z., Ha, S.-Y., and Xue, X.: Emergent phenomena in an ensemble of Cucker-Smale particles under joint rooted leadership, *Math. Mod. Meth. Appl. Sci.* **24**, 1389–1419 (2014).
49. Leonard, N. E., Paley, D. A., Lekien, F., Sepulchre, R., Fratantoni, D.M. and Davis, R. E.: Collective motion, sensor networks and ocean sampling, *Proc. IEEE* **95**, 48–74 (2007).
50. Li, Z. and Xue, X.: Cucker-Smale flocking under rooted leadership with fixed and switching topologies, *SIAM J. Appl. Math.* **70**, 3156–3174 (2010).
51. Motsch, S. and Tadmor, E.: Heterophilious dynamics enhances consensus, *SIAM Rev.* **56**, 577–621 (2014).
52. Motsch, S. and Tadmor, E.: A new model for self-organized dynamics and its flocking behavior, *J. Stat. Phys.* **144**, 923–947 (2011).
53. Nagy, M., Ákos, Z., Biro, D., and Vicsek, T.: Hierarchical group dynamics in pigeon flocks, *Nature* **464**, 890–893 (2010).
54. O'Rourke, P.: Collective drop effects on vaporising liquid sprays, Ph. D. Thesis, Princeton University, Princeton, NJ, 1981.
55. Paley, D.A., Leonard, N. E., Sepulchre, R., Grunbaum, D. and Parrish, J. K.: Oscillator models and collective motion, *IEEE Control Systems* **27**, 89–105 (2007).

56. Park, J., Kim, H., and Ha, S.-Y.: Cucker-Smale flocking with inter-particle bonding forces, *IEEE Tran. Automatic Control* **55**, 2617–2623 (2010).
57. Perea, L., Gómez, G., and Elosegui, P.: Extension of the Cucker-Smale control law to space flight formation, *J. Guidance, Control and Dynamics* **32**, 526–536 (2009).
58. Reynolds, C. W.: Flocks, herds and schools: A distributed behavioral model, *Proceeding SIGGRAPH 87 Proceedings of the 14th annual conference on Computer graphics and interactive techniques* 25–34 (1987).
59. Ranz, W. and Marshall, W.: Evaporation from drops, *Chem. Eng. Prog.* **48**, 141–180 (1952).
60. Shen, J.: Cucker-Smale Flocking under Hierarchical Leadership, *SIAM J. Appl. Math.* **68**, 694–719 (2007).
61. Spannenberg, A. and Galvin, K. P.: Continuous differential sedimentation of a binary suspension, *Chem. Eng. Aust.* **21**, 7–11 (1996).
62. Tadmor, E. and Tan, C.: Critical thresholds in flocking hydrodynamics with nonlocal alignment, *Proc. Royal Soc. A*, 372:20130401 (2014).
63. Toner, J. and Tu, Y.: Flocks, herds, and Schools: A quantitative theory of flocking, *Physical Review E* **58**, 4828–4858 (1998).
64. Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen I., and Shochet O.: Novel type of phase transition in a system of self-driven particles, *Phys. Rev. Lett.* **75**, 1226–1229 (1995).
65. Vinkovic, I., Aguirre, C., Simoëns S., and Gorokhovski, M.: Large eddy simulation of droplet dispersion for inhomogeneous turbulent wall flow, *Int. J. Multiph. Flow* **32**, 344–364 (2006).
66. Williams, F. A.: Spray combustion and atomization, *Phys. fluids* **1**, 541–555 (1958).
67. Xue, X., and Guo, L.: A kind of nonnegative matrices and its application on the stability of discrete dynamical systems, *J. Math. Anal. Appl.* **331**, 1113–1121 (2007).
68. Xue, X., and Li, Z.: Asymptotic stability analysis of a kind of switched positive linear discrete systems, *IEEE Trans. Autom. Control* **55**, 2198–2203 (2010).

Follow-the-Leader Approximations of Macroscopic Models for Vehicular and Pedestrian Flows

M. Di Francesco, S. Fagioli, M.D. Rosini and G. Russo

Abstract We review the recent results and present new ones on a deterministic *follow-the-leader* particle approximation of first-and second-order models for traffic flow and pedestrian movements. We start by constructing the particle scheme for the first-order Lighthill–Whitham–Richards (LWR) model for traffic flow. The approximation is performed by a set of ODEs following the position of discretized vehicles seen as moving particles. The convergence of the scheme in the many particle limit toward the unique entropy solution of the LWR equation is proven in the case of the Cauchy problem on the real line. We then extend our approach to the initial–boundary value problem (IBVP) with time-varying Dirichlet data on a bounded interval. In this case, we prove that our scheme is convergent strongly in \mathbf{L}^1 up to a subsequence. We then review extensions of this approach to the Hughes model for pedestrian movements and to the second-order Aw–Rascle–Zhang (ARZ) model for vehicular traffic. Finally, we complement our results with numerical simulations. In particular, the simulations performed on the IBVP and the ARZ model suggest the consistency of the corresponding schemes, which is easy to prove rigorously in some simple cases.

M. Di Francesco (✉) · S. Fagioli
DISIM, Università degli Studi dell’Aquila, via Vetoio 1 (Coppito),
67100 L’Aquila, AQ, Italy
e-mail: marco.difrancesco@univaq.it

S. Fagioli
e-mail: simone.fagioli@dm.univaq.it

M.D. Rosini
Instytut Matematyki, Uniwersytet Marii Curie-Skłodowskiej,
plac Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland
e-mail: mrosini@umcs.lublin.pl

G. Russo
Dipartimento di Matematica ed Informatica, Università di Catania,
Viale Andrea Doria 6, 95125 Catania, Italy
e-mail: russo@dmi.unict.it

1 Introduction

The modeling of vehicular traffic flow can be considered as one of the most important challenges of applied mathematics within the last seventy years. Among its several repercussions on real-world applications, we mention, e.g., the development of smart traffic management systems for integrated applications of communications, control, and information processing technologies to the whole transport system. Other important resultant benefits are the implementation of complex problem-solving in traffic management and the addressing of practical problems such as reducing congestion and related costs. These goals can be achieved by optimizing the use of transport resources and infrastructures of the transport system as a whole, by bringing more efficiency in terms of traffic fluidity, and by providing procedures for system stabilization.

Several analytical models for vehicular traffic have been developed in the last decades. In the first instance, they are classified into two main classes: microscopic models—taking into account each single vehicle—and macroscopic ones—dealing with averaged quantities. We refer to [12, 13, 61, 64] for a survey of the most commonly used models currently available in the literature.

Recently, the availability of online data allows the implementation of real-time strategies aiming at avoiding (or mitigating) congested traffic. To address this task, the development and the application of analytical models that are easy-to-use and with a high performance in terms of time and reliability are essential requirements. In this sense, opposed to direct numerical ‘individual-based’ simulations of a large number of interacting vehicles—as typical when dealing with microscopic models—many researchers recommend using macroscopic models for traffic flow. The main advantages of the macroscopic approach with respect to the microscopic one are as follows:

- The model is completely evolutive and is able to rapidly describe traffic situations at every time;
- The resulting description of queues evolution and of traveling times is accurate as the position of shock waves can be exactly computed and corresponds to queue tails;
- The macroscopic theory helps developing efficient numerical schemes suitable to describe very large number of vehicles;
- The model can be easily calibrated, validated, and implemented as the number of parameters is low;
- The theory allows to state and possibly solve optimal management problems.

The macroscopic variables are the density ρ (number of vehicles per unit length of the road), the velocity v (space covered per unit time by the vehicles), and the flow f (number of vehicles per unit time). Clearly, the macroscopic variables are in general functions of time $t > 0$ and space $x \in \mathbb{R}$. By definition

$$f = \rho v. \quad (1)$$

Moreover, the conservation of the number of vehicles along a road with neither entrances nor exits is expressed by the one-dimensional scalar conservation law [18]

$$\rho_t + f_x = 0. \quad (2)$$

The systems (1)–(2) has three unknown variables. Hence, a further condition has to be imposed. There are two main approaches to do it: first- and second-order models. First-order models introduce a further explicit expression of one of the three unknown variables in terms of the remaining two. The prototype of first-order model is the Lighthill, Whitham [53], and Richards [63] (LWR) model. The basic assumption of LWR is that the velocity of any driver depends on the density alone

$$v = \mathcal{V}(\rho),$$

where $\mathcal{V} \in \mathbf{C}^1([0, \rho_{\max}]; [0, v_{\max}])$ is non-increasing, with $\mathcal{V}(\rho_{\max}) = 0$ and $\mathcal{V}(0) = v_{\max} > 0$, where $\rho_{\max} > 0$ is the maximal density corresponding to the ‘bumper-to-bumper’ situation, and v_{\max} is the maximal speed corresponding to the free road. As a result, the LWR model is given by the scalar conservation law

$$\rho_t + [\rho \mathcal{V}(\rho)]_x = 0. \quad (3)$$

Second-order macroscopic models close the systems (1) and (2) by adding a further conservation law. The most celebrated second-order macroscopic model is the Aw, Rascle [9], and Zhang [69] (ARZ) model. Away from the vacuum state $\rho = 0$, the ARZ model writes

$$\rho_t + [\rho v]_x = 0, \quad [\rho (v + p(\rho))]_t + [\rho (v + p(\rho)) v]_x = 0, \quad (4)$$

where the function $p(\rho)$ is introduced to take into account drivers’ reactions to the state of traffic in front of them.

The main drawback of the LWR model is the unrealistic behavior of the drivers adjusting instantaneously their velocities according to the densities they are experiencing. Moreover, the graph of a map $\rho \mapsto [\rho \mathcal{V}(\rho)]$ cannot represent the cloud of points in the (ρ, f) -plane obtained by empirical measurements. The ARZ model avoids these drawbacks of the LWR model. However, the system (4) degenerates into just one equation at the vacuum state $\rho = 0$. In particular, the solutions to the ARZ model do not depend continuously on the initial data in any neighborhood of $\rho = 0$.

We point out that (1) and (2) are the only accurate physical laws in vehicular traffic theory. All other equations result from coarse approximations of empirical observations. However, as the dynamics of any living system are influenced by psychological effects, nobody would expect that traffic models could reach an accuracy comparable to that attained in other domains of science, such as thermodynamics or Newtonian physics. Nevertheless, they can have sufficient descriptive power for

the specific application-driven purpose, and they can help understanding non-trivial phenomena of vehicular traffic.

The use of macroscopic models relies on the *continuum assumption*, namely on the assumption that the medium is indefinitely divisible without changing its physical nature. This assumption is not justifiable in the context of vehicular traffic, but is accepted as a technical hypothesis. In order to make more clear the continuum hypothesis, the study of the micro-to-continuum limit for first- and second-order models has been proposed in [24, 26] and [8, 14], respectively. Our goal is to address said discrete-to-continuum limit in a rigorous analytic form, for both first- and second-order models, by proving that the macroscopic models can be *solved* as a *many particle limit* of discrete (microscopic) ODE-based models.

We sketch here our approach for the LWR model (3), described in detail in Section 2.1. Fix an initial density $\bar{\rho}$. Let $L \doteq \|\bar{\rho}\|_{L^1(\mathbb{R})}$ be the total space occupied by the all vehicles (i.e., the total mass in a ‘continuum PDEs’ language). For a given positive $n \in \mathbb{N}$, we split $\bar{\rho}$ into n platoons of ‘possibly fractional’ vehicles, each one of equal length $\ell_n \doteq L/n$, with the endpoints of each platoon positioned at $\bar{x}_i \in \mathbb{R}$, $i = 0, \dots, n$. The points \bar{x}_i are taken as initial condition to the microscopic follow-the-leader (FTL) model for vehicular traffic

$$\begin{cases} \dot{x}_i(t) = \mathcal{V} \left(\frac{\ell_n}{x_{i+1}(t) - x_i(t)} \right), & i \in \{0, \dots, n-1\}, \\ \dot{x}_n(t) = v_{\max}. \end{cases} \quad (5)$$

The points $x_i(t)$ are interpreted as moving particles along the real line \mathbb{R} . In Lemma 1, below, we prove that no collisions occur between the particles, as the distance between two consecutive particles is bounded from below by ℓ_n/ρ_{\max} for all times. We then consider the *discrete density*

$$\rho^n(t, x) \doteq \sum_{i=0}^{n-1} R_i^n(t) \mathbf{1}_{[x_i(t), x_{i+1}(t))}, \quad R_i^n(t) \doteq \frac{\ell_n}{x_{i+1}(t) - x_i(t)},$$

and prove that (up to a subsequence) its limit as $n \rightarrow \infty$ is the entropy solution to the LWR model (3) in the Oleinik–Kruzhkov sense [52, 59]. The convergence of the particle scheme (5) toward (3) is proven rigorously in [33], see also the improved results in [30]. We refer to Section 2 for the details.

The result in [30, 33] can be interpreted as a *particle method* for the one-dimensional scalar conservation law (3), which can be applied in the context of numerics. Particle methods feature a long-standing history as a numerical method for transport equations (see, e.g., [57] and the references therein). Moreover, several effective numerical schemes for nonlinear conservation laws are proposed in the literature. We mention the pioneering work of Glimm [41] for systems, and the wave-front tracking (WFT) algorithm proposed by Dafermos in [27] and improved later on by Di Perna [34] and Bressan [17], see also [48] and the references therein for more details. Our approach differs from most of the numerical methods for scalar

conservation laws in that it interprets the microscopic limit as a *mean field limit of a system of interacting particles with nearest neighbor-type interaction*, in the spirit of interacting particles systems in probability, kinetic theory, statistical mechanics, and mathematical biology (see, e.g., [35, 56, 60]). We stress in particular the fundamental role of many particle exclusion processes in probability, a subject which has been extensively studied in a vast literature in the past decades (see, e.g., [38, 39, 51] and the references therein). It is worth recalling at this stage that Lions, Perthame, and Tadmor proved in [54] that nonlinear conservation laws can also be solved via kinetic approximation.

Unlike in most of the aforementioned articles, our approach should be regarded as a *deterministic particle approximation* to the target PDE's. A pioneering result is the one by Russo [65], which applies to the linear diffusion equation with the diffusion operator replaced by a nearest neighbor interaction term (see also later generalizations in [44, 55]). Our approach can be considered in the spirit of [65], applied to scalar conservation laws. We also mention the paper by Brenier and Grenier [16], which provides a particle approximation of the pressureless Euler system.

Our approach follows essentially the same strategy in the uniform estimates adopted for the WFT algorithm, except that a lighter notion of time continuity is needed involving (a scaled version of) the *1-Wasserstein distance* (see Section 1.1 or [4, 68] for more details). A major advantage in using the Wasserstein distance relies on its identification with the \mathbf{L}^1 -topology in the space of *pseudo-inverses of cumulative distributions*. Such an identification allows to recover formally the ODE system (5) as the most natural way to approximate (3) via Lagrangian particles. We briefly sketch it here. Let ρ be the solution to (3), and let

$$F(t, x) \doteq \int_{-\infty}^x \rho(t, x) dx \in [0, L],$$

be its primitive. The pseudo-inverse variable $X(t, z) \doteq \inf \{x \in \mathbb{R}: F(x) > z\}$, $z \in [0, L]$, formally satisfies the *Lagrangian PDE*

$$X_t(t, z) = \mathcal{V}(X_z(t, z)^{-1}).$$

Now, if we replace the above z -derivative by a forward finite difference

$$X_z \approx \frac{X(t, z + \ell_n) - X(t, z)}{\ell_n},$$

and assume that X is piecewise constant on intervals of length ℓ_n , the ODE system (5) is immediately recovered, with the structure

$$X(t, z) = \sum_i x_i(t) \chi_{[i\ell_n, (i+1)\ell_n)}(z).$$

The use of pseudo-inverse variables and Wasserstein distances in the framework of scalar conservation laws is not totally new (see, e.g., [15, 20]). As far as the LWR model is concerned, in [58] a simplified version of the LWR model is derived by introducing as new variable the cumulative number of vehicles passing through a location x at time t (see also [7, 28]).

A natural question concerning the particle approximation procedure described above is whether or not it can be applied to recover the solution to the IBVP with *Dirichlet boundary condition*

$$\begin{cases} \rho_t + f(\rho)_x = 0, & x \in (0, 1), t \in (0, T), \\ \rho(0, x) = \bar{\rho}(x), & x \in (0, 1), \\ \rho(t, 0) = \bar{\rho}_0(t), & t \in (0, T), \\ \rho(t, 1) = \bar{\rho}_1(t), & t \in (0, T). \end{cases} \quad (6)$$

Such a question is addressed for the first time in the present work. Due to the propagation of the initial and boundary conditions along characteristic lines, it is well known that a concept of Dirichlet condition for a nonlinear conservation law has to be formulated in a set-valued sense. The first rigorous definition of entropy solution in this context was provided in [10], in which existence and uniqueness were proven in the scalar multidimensional case. In the one-dimensional case, a more intuitive notion of entropy solution was provided in [36], where the authors proved that at least in the scalar case, the trace of the solution at the boundary is obtained by solving a Riemann problem within the trace itself and the boundary datum.

The substantial mismatch between Lagrangian and Eulerian speeds of propagation suggests that prescribing the behavior of the particle system (5) near the boundary should not involve characteristic speeds. Inspired by the extremely simple structure of the FTL system (5), the boundary dynamics should follow a very natural process, possibly reminiscent of empirical observation in real contexts (e.g., a toll gate). At the same time, such a dynamics should be able to capture the notion of entropy solution for the limiting IBVP for a large number of particles. Our choice for the definition of the scheme in this case is pretty natural. We sketch it here in the simple case of constant boundary conditions $\rho(t, 0^+) = \bar{\rho}_0$, $\rho(t, 1^-) = \bar{\rho}_1$.

Initially, $n + N + 1$ particles of mass $\ell_n \doteq n^{-1} \|\bar{\rho}\|_{L^1(0,1)}$ are set in $\bar{x}_{-N}, \dots, \bar{x}_n$ with $\bar{x}_{-N} < \dots < \bar{x}_{-1} < \bar{x}_0 \doteq 0 < \bar{x}_1 < \dots < \bar{x}_{n-1} < \bar{x}_n = 1$. The *entering* condition is set by requiring that $\bar{x}_i \doteq i \ell_n / \bar{\rho}_0$, $i \in \{-N, \dots, -1\}$, so that the queuing particles in $x < 0$ are equidistant and matching the boundary datum $\bar{\rho}_0$. The *exit* condition is set by requiring that $\dot{x}_n = \mathcal{V}(\bar{\rho}_1)$. We then let evolve the particles according to the corresponding version of the FTL scheme (5). After some time, some of the queuing vehicles will enter the domain $[0, 1]$ and some particle will leave it. In general, in a finite time, the distances between the particles in $x < 0$ will not match the boundary datum $\bar{\rho}_0$, as well as the leftmost particle in $x \geq 1$ will not necessarily move with velocity $\mathcal{V}(\bar{\rho}_1)$. For this reason, we introduce a sufficiently small time step $\tau > 0$ and, at each time $t = k \tau$, $k \in \mathbb{N}$, we rearrange the particles in both $x < 0$ and $x > 1$ so that the resulting densities match the corresponding boundary data, while

on each time interval $[k\tau, (k+1)\tau)$, $k \in \mathbb{N}$, we let the particles evolve according to the corresponding version of FTL scheme (5), with $\dot{x}_n = \mathcal{V}(\bar{\rho}_1)$. Let us underline that the number N (which depends on n) should be prescribed initially depending on the final time T , in a way that some of the queuing vehicles are still left in $x < 0$ at time $t = T$.

In order to extend our approach to time-varying boundary data, we discretise the boundary conditions with respect to time via a time step τ , solve the particle system in each time interval with constant boundary data, and then rearrange the particles outside the domain according to the boundary condition at the next time step. We defer to Section 3.1 for more details. We remark that in the case of constant boundary conditions for the continuum equation (3), the rearrangement of the boundary datum at each time step τ is not necessary as long as no waves hit the boundary from the interior of the domain. Such a situation also holds in our particle approximation, as we shall see in the simulations in Subsection 6.2.

We prove rigorously in Section 3 that the above particle scheme converges strongly in L^1 to a limiting density ρ as $n \rightarrow +\infty$ and $\tau \rightarrow 0$. Such a result does not require any condition on how fast (or slow) n should tend to infinity with respect to τ tending to zero. The consistency of the scheme is provided in simple cases, i.e., either constant initial data or boundary conditions yielding outgoing characteristic speeds at the boundary. As we explain below in Section 3, the definition of our approximating scheme is reminiscent of the notion of entropy solution provided in [36], see Definition 2, in which the trace of the solution ρ to (3) at the boundary is required to match the solution to a suitable Riemann problem. Our scheme actually prescribes a constant datum outside the domain at each time step, in a way to produce the approximation to a Riemann problem near the boundary. The simulations we provide in Subsection 6.2 support our conjecture that our scheme is consistent with the notion of entropy solution in the sense of [10, 36].

The deterministic particle approach started in [33] has seen significant extensions to similar models. A first one has been performed in [29] on the ARZ model (4). Despite the second-order nature of ARZ, the strategy developed in [33] for the first-order LWR model (3) applies also in this case. This reveals that the multi-species nature of the ARZ model is quite relevant in the dynamics. Our rigorous results only deal with the convergence toward a weak solution. The problem of the uniqueness of entropy solutions for the ARZ model is quite a hard task. For this reason, we do not address here the consistency of our scheme. Let us point out that our approach for the ARZ system deeply differs from the one proposed in [8], which essentially works away from the vacuum state and is implemented via a time discretization and suitable space–time scaling. Our result in [29] works near the vacuum state, and no scaling is performed. Unlike the previous numerical attempts (e.g., [22]), our method is conservative and is able to cope with the vacuum. We briefly recall the result of [29] in Section 5 below.

Another extension of our particle approach has been performed in [31] on a one-dimensional version of the *Hughes model* [49] for pedestrian movements, (see (30) below). In this model, the movement of a dense human crowd is modeled via a ‘thinking fluid’ approach in which the crowd is modeled as a continuum medium, with

Eulerian velocity computed via a non-local constitutive law of the overall distribution of pedestrians. Such a non-local dependence is encoded in the *weighted* distance function ϕ , computed at a quasi-equilibrium regime via a nonlinear *running cost* function $c(\rho)$. The function ϕ may be interpreted as an estimated exit time for a given distribution of pedestrians. We refer to [11, 64] and the references therein for the mathematical modeling of human crowds, and to [2, 3, 19, 21, 32, 37, 42, 49] for the rigorous analytical results and numerical simulations available in the literature on the Hughes model.

A fully satisfactory existence theory for the Hughes model is still missing. A mathematical theory in this setting was first addressed in [32], in which the eikonal equation was replaced by two regularized versions involving a Laplacian term. A rigorous mathematical treatment of the Riemann problems for the Hughes model without any regularization was performed independently in [2] and [37]. Said result led the basis to tackle the existence theory via a WFT strategy. As in the paper [3], we prove in [31] the existence of entropy solutions when the initial condition yields the formation of two distinct groups of pedestrians moving toward the two exits, with the emergence of a vacuum region in between, persisting until the total evacuation of the domain. However, differently from [3] where the WFT method is applied, in [31] we develop a FTL particle approximation, taking advantage of the fact that our assumptions ensure that the Hughes model can be formulated as a two-sided LWR equations. We refer to [31] and to Section 4.1 below for the precise formulation of the particle scheme. As a result, we prove that the particle scheme converges under (essentially) the same conditions for which an existence result for entropy solutions is available in the literature (with the results in [3] in mind).

This chapter is structured as follows:

- In Section 2, we review the results in [33] and later improvements in [30] about the convergence of the FTL scheme (5) toward entropy solutions to the LWR equation (3). The main result is stated in Theorem 3.
- In Section 3, we prove our new result concerning the convergence of the FTL scheme for the IBV problem (6). The strong convergence of the scheme is proven in Theorem 4. The consistency of the scheme in some special cases is proven in Theorem 5.
- In Section 4, we review the results in [31] on the particle approximation of the Hughes model (30), with the main result stated in Theorem 7.
- In Section 5, we review the results in [29] on the ARZ model (4). The main result is stated in Theorem 8.
- In Section 6, we collect all the numerical simulations performed for the particle methods introduced in all the aforementioned models. In particular, we present new simulations regarding the IBV problem (6) in Subsection 6.2

In the next subsection, we recall the basic results on the Wasserstein distance that are used in this chapter.

1.1 The Wasserstein Distances

We collect here the main concepts about one-dimensional Wasserstein distances (see [68] for further details). As already mentioned, we deal with probability densities with constant mass in time and we need to evaluate their distances at different times in the Wasserstein sense.

For a fixed mass $L > 0$, we consider the space

$$\mathcal{M}_L \doteq \{\mu \text{ Radon measure on } \mathbb{R} \text{ with compact support: } \mu \geq 0 \text{ and } \mu(\mathbb{R}) = L\}.$$

For a given $\mu \in \mathcal{M}_L$, we introduce the pseudo-inverse variable $X_\mu \in \mathbf{L}^1([0, L]; \mathbb{R})$ as

$$X_\mu(z) \doteq \inf\{x \in \mathbb{R}: \mu((-\infty, x]) > z\}. \quad (7)$$

Clearly, X_μ is non-decreasing on $[0, L]$ and locally constant on ‘mass intervals’ on which μ is concentrated. X_μ may have (increasing) jumps if the support of μ is not connected. By abuse of notation, in case $\mu = \rho \mathcal{L}_1$ is absolutely continuous with respect to the Lebesgue measure, we denote its pseudo-inverse variable by X_ρ .

For $L = 1$, the one-dimensional 1-Wasserstein distance between $\rho_1, \rho_2 \in \mathcal{M}_1$ (defined in terms of optimal plans in the Monge-Kantorovich problem, see, e.g., [68]) can be defined as

$$W_1(\rho_1, \rho_2) \doteq \|X_{\rho_1} - X_{\rho_2}\|_{\mathbf{L}^1([0, 1]; \mathbb{R})}.$$

We introduce the scaled 1-Wasserstein distance between $\rho_1, \rho_2 \in \mathcal{M}_L$ as

$$W_{L,1}(\rho_1, \rho_2) \doteq \|X_{\rho_1} - X_{\rho_2}\|_{\mathbf{L}^1([0, L]; \mathbb{R})}. \quad (8)$$

Indeed, a straightforward computation yields $W_{L,1}(\rho_1, \rho_2) = L W_1(L^{-1} \rho_1, L^{-1} \rho_2)$. The distance $W_{L,1}$ inherits all the topological properties of the 1-Wasserstein distance for probability measures. In particular, a sequence $(\rho_n)_n$ in \mathcal{M}_L converges to $\rho \in \mathcal{M}_L$ in $W_{L,1}$ if and only if for any $\varphi \in \mathbf{C}^0(\mathbb{R}; \mathbb{R})$ growing at most linearly at infinity

$$\lim_{n \rightarrow +\infty} \int_{\mathbb{R}} \varphi(x) d\rho_n(x) = \int_{\mathbb{R}} \varphi(x) d\rho(x).$$

We now state a technical result which will serve in the sequel of the chapter.

Theorem 1 (Generalized Aubin-Lions lemma) *Assume $v : [0, +\infty) \rightarrow [0, +\infty)$ is a continuous and strictly monotone function. Let $T, L > 0$, $a, b \in \mathbb{R}$ be fixed with $a < b$. Let $(\rho^n)_n$ be a sequence in $\mathbf{L}^\infty((0, T); \mathbf{L}^1(\mathbb{R}))$ with $\rho^n(t, \cdot) \geq 0$ and $\|\rho^n(t, \cdot)\|_{\mathbf{L}^1(\mathbb{R})} = L$ for all $n \in \mathbb{N}$ and $t \in [0, T]$. Assume further that*

$$\sup_{n \in \mathbb{N}} \left[\int_0^T \left[\|v(\rho^n(t, \cdot))\|_{\mathbf{L}^1([a, b])} + \text{TV}(v(\rho^n(t, \cdot)); [a, b]) \right] dt \right] < +\infty, \quad (\text{H1})$$

$$\lim_{h \downarrow 0} \left[\sup_{n \in \mathbb{N}} \left[\int_0^{T-h} W_{L,1}(\rho^n(t+h, \cdot), \rho^n(t, \cdot)) dt \right] \right] = 0. \quad (\text{H2})$$

Then, $(\rho^n)_n$ is strongly relatively compact in $\mathbf{L}^1([0, T] \times [a, b])$.

We refer to the Appendix of [30] for the proof of Theorem 1. We will sometimes consider the following condition:

$$\begin{aligned} & \text{There exists a constant } C > 0 \text{ independent of } n \text{ such that} \\ & W_{L,1}(\rho^n(t, \cdot), \rho^n(s, \cdot)) \leq C |t - s| \text{ for all } s, t \in (0, T). \end{aligned} \quad (\text{H2}')$$

We point out that (H2') implies (H2) and that it is assumed in both [30, Theorem 3.5] and [33, Theorem 3.2].

2 The LWR Model

In this section, we review the results obtained in [33], later improved in [30], on the Cauchy problem for the LWR model (3)

$$\begin{cases} \rho_t + f(\rho)_x = 0, & (t, x) \in \mathbb{R}_+ \times \mathbb{R}, \\ \rho(0, x) = \bar{\rho}(x), & x \in \mathbb{R}, \end{cases} \quad (9)$$

where $f(\rho) \doteq \rho v(\rho)$. If $\rho_{\max} > 0$ is the maximal density corresponding to the situation in which the vehicles are bumper to bumper, and v_{\max} is the maximal speed corresponding to the free road, then the initial datum $\bar{\rho}$ and the velocity map v are assumed to satisfy the following conditions:

$$\bar{\rho} \in \mathbf{L}^\infty \cap \mathbf{L}^1(\mathbb{R}; [0, \rho_{\max}]), \quad (\text{I1})$$

$$v \in \mathbf{C}^1([0, \rho_{\max}]; [0, v_{\max}]), \quad v' < 0, \quad v(0) = v_{\max}, \quad v(\rho_{\max}) = 0. \quad (\text{V1})$$

In some cases we require also one of the following conditions:

$$\bar{\rho} \in \mathbf{BV}(\mathbb{R}; [0, \rho_{\max}]), \quad (\text{I2})$$

$$[0, \rho_{\max}] \ni \rho \mapsto \rho v'(\rho) \in \overline{\mathbb{R}}_- \text{ is non-increasing.} \quad (\text{V2})$$

Example 1 (Examples of velocities in vehicular traffic). The prototype for the velocity in vehicular traffic $v_{GS}(\rho) \doteq v_{\max} \left(1 - \frac{\rho}{\rho_{\max}}\right)$ by Greenshields [46] clearly satisfies the assumptions (V1), (V2). The same holds for the Pipes–Munjal velocity [62]

$$v_{PM}(\rho) \doteq v_{\max} \left[1 - \left(\frac{\rho}{\rho_{\max}} \right)^{\alpha} \right], \quad \alpha > 0,$$

in which the concavity of the flux $\rho v_{PM}(\rho)$ degenerates at $\rho = 0$. Further examples of speed–density relations that satisfy (V1), (V2) are

$$\begin{aligned} v_{GB}(\rho) &\doteq v_{\max} \left[\log \left(\frac{\rho_{\max} + \alpha}{\alpha} \right) \right]^{-1} \log \left(\frac{\rho_{\max} + \alpha}{\rho + \alpha} \right), \quad \alpha > 0, \\ v_U(\rho) &\doteq v_{\max} [1 - e^{-\rho_{\max}}]^{-1} [e^{-\rho} - e^{-\rho_{\max}}], \end{aligned}$$

that result from a slight modification of that ones proposed by Greenberg [45] and Underwood [67] respectively.

Definition 1 Assume (I1) and (V1). We say that $\rho \in \mathbf{L}^\infty(\mathbb{R}_+ \times \mathbb{R})$ is an *entropy solution* to the Cauchy problem (9) if $\rho(t, \cdot) \rightarrow \bar{\rho}$ in the weak* \mathbf{L}^∞ sense as $t \downarrow 0$ and

$$\iint_{\mathbb{R}_+ \times \mathbb{R}} \left[|\rho(t, x) - k| \varphi_t(t, x) + \text{sign}(\rho(t, x) - k) [f(\rho(t, x)) - f(k)] \varphi_x(t, x) \right] dx dt \geq 0$$

for all $\varphi \in \mathbf{C}_c^\infty((0, +\infty) \times \mathbb{R})$ with $\varphi \geq 0$ and for all $k \geq 0$.

We point out that the above definition is slightly weaker than the definition in [52]. The next theorem collects the uniqueness result in [52] and its variant in [23].

Theorem 2 ([23, 52]) Assume (I1) and (V1). Then there exists a unique entropy solution to the Cauchy problem (9) in the sense of Definition 1.

2.1 The Follow-the-Leader Scheme and Main Result

We now introduce rigorously our FTL approximation scheme for (9). Assume (I1) and (V1). Let

$$L \doteq \|\bar{\rho}\|_{\mathbf{L}^1(\mathbb{R})}, \quad R \doteq \|\bar{\rho}\|_{\mathbf{L}^\infty(\mathbb{R})}.$$

Fix $n \in \mathbb{N}$ sufficiently large. Let $\ell_n \doteq L/n$ and $\bar{x}_1^n, \dots, \bar{x}_{n-1}^n$ be defined recursively by

$$\begin{cases} \bar{x}_1^n \doteq \sup \left\{ x \in \mathbb{R}: \int_{-\infty}^x \bar{\rho}(x) dx < \ell_n \right\}, \\ \bar{x}_i^n \doteq \sup \left\{ x \in \mathbb{R}: \int_{\bar{x}_{i-1}^n}^x \bar{\rho}(x) dx < \ell_n \right\}, \quad i \in \{2, \dots, n-1\}. \end{cases}$$

It follows that $\bar{x}_1^n < \bar{x}_2^n < \dots < \bar{x}_{n-1}^n$ and

$$\int_{-\infty}^{\bar{x}_1^n} \bar{\rho}(x) dx = \int_{\bar{x}_{i-1}^n}^{\bar{x}_i^n} \bar{\rho}(x) dx = \int_{\bar{x}_{i-1}^n}^{+\infty} \bar{\rho}(x) dx = \ell_n \leq (\bar{x}_i^n - \bar{x}_{i-1}^n) R, \quad i \in \{2, \dots, n-1\}. \quad (10)$$

We let the $(n-1)$ particles defined above evolve according to the FTL system

$$\begin{cases} \dot{x}_i^n(t) = v(R_i^n(t)), & i \in \{1, \dots, n-2\}, \\ \dot{x}_{n-1}^n(t) = v_{\max}, \\ x_i^n(0) = \bar{x}_i^n, & i \in \{1, \dots, n-1\}, \end{cases} \quad R_i^n(t) \doteq \frac{\ell_n}{x_{i+1}^n(t) - x_i^n(t)}. \quad (11)$$

Lemma 1 (Discrete maximum principle [33, Lemma 1]). Assume (I1) and (V1). Then, the solution $(x_i^n)_{i=1}^{n-1}$ to (11) satisfies for any $t \geq 0$

$$x_{i+1}^n(t) - x_i^n(t) \geq \ell_n/R, \quad i \in \{1, \dots, n-2\}.$$

The above lemma ensures that the particles strictly preserve their initial order. Hence, the solution $(x_i^n)_{i=1}^{n-1}$ to (11) is well defined.

We introduce two *artificial particles* x_0^n and x_n^n as follows:

$$x_0^n(t) \doteq 2x_1^n(t) - x_2^n(t), \quad x_n^n(t) \doteq 2x_{n-1}^n(t) - x_{n-2}^n(t), \quad (12)$$

and let $R_0^n \doteq R_1^n$ and $R_{n-1}^n \doteq R_{n-2}^n$. We then set

$$\rho^n(t, x) \doteq \sum_{i=0}^{n-1} R_i^n(t) \mathbf{1}_{[x_i^n(t), x_{i+1}^n(t))}(x) = \sum_{i=0}^{n-1} \frac{\ell_n}{x_{i+1}^n(t) - x_i^n(t)} \mathbf{1}_{[x_i^n(t), x_{i+1}^n(t))}(x). \quad (13)$$

We notice that $\|\rho^n(t, \cdot)\|_{L^1(\mathbb{R})} = L$, $\|\rho^n(t, \cdot)\|_{L^\infty(\mathbb{R})} \leq R$ and that $\rho^n(t, \cdot)$ is compactly supported for all $t \geq 0$. For future use, we compute

$$\begin{cases} \dot{R}_i^n(t) = -\frac{R_i^n(t)^2}{\ell_n} [v(R_{i+1}^n(t)) - v(R_i^n(t))], & i \in \{1, \dots, n-3\}, \\ \dot{R}_{n-2}^n(t) = -\frac{R_{n-2}^n(t)^2}{\ell_n} [v_{\max} - v(R_{n-2}^n(t))]. \end{cases} \quad (14)$$

Remark 1 In case $\text{supp}[\bar{\rho}]$ is bounded either from above or from below, it is possible to improve the above construction. In the former case, the particle x_n^n can be set on $\max\{\text{supp}[\bar{\rho}]\}$ initially and let evolve with maximum speed v_{\max} , and the preceding particle x_{n-1}^n let evolve according to $\dot{x}_{n-1}^n(t) = v(\ell_n/(x_n^n(t) - x_{n-1}^n(t)))$. In the latter case, the particle x_0^n can be set on $\min\{\text{supp}[\bar{\rho}]\}$ initially and let evolve according to $\dot{x}_0^n(t) = v(\ell_n/(x_1^n(t) - x_0^n(t)))$. In [33], both these conditions are required for the initial datum and such construction is applied.

The main result of [30, 33] reads as follows.

Theorem 3 ([30, Theorem 2.3], [33, Theorem 3]). Assume (I1) and (V1). Moreover, assume at least one of the two conditions (I2) and (V2). Then, $(\rho^n)_n$ converges (up to a

subsequence) a.e. and in $\mathbf{L}_{\text{loc}}^1$ on $\mathbb{R}_+ \times \mathbb{R}$ to the unique entropy solution to the Cauchy problem (9) in the sense of Definition 1.

We sketch the proof of Theorem 3 in the next two subsections. For simplicity, we assume that $\bar{\rho}$ is compactly supported and apply the corresponding construction explained in Remark 1.

2.2 Estimates

The result in Lemma 1 ensures that $\|\rho^n(t, \cdot)\|_{\mathbf{L}^\infty(\mathbb{R})} \leq R \doteq \|\bar{\rho}\|_{\mathbf{L}^\infty(\mathbb{R})}$ for all $t \geq 0$. As usual in the context of scalar conservation laws, a uniform control of the **BV** norm is necessary in order to gain enough compactness of the approximating scheme. We achieve compactness in two distinct ways. The first one is a uniform **BV** contraction property for $(\rho^n)_n$, and it obviously requires (I2).

Proposition 1 *Assume (I1), (I2) and (V1). Then, the discretized density ρ^n defined in (13) satisfies for any $t \geq 0$*

$$\text{TV}[\rho^n(t, \cdot)] \leq \text{TV}[\rho^n(0, \cdot)] \leq \text{TV}[\bar{\rho}].$$

Proof By (14) and (V1) we have that

$$\begin{aligned} & \frac{d}{dt} \text{TV}[\rho^n(t, \cdot)] = \\ &= [1 + \text{sign}(R_1(t) - R_2(t))] \dot{R}_1(t) + [1 - \text{sign}(R_{n-3}(t) - R_{n-2}(t))] \dot{R}_{n-2}(t) \\ &+ \sum_{i=2}^{n-3} [\text{sign}(R_i(t) - R_{i+1}(t)) - \text{sign}(R_{i-1}(t) - R_i(t))] \dot{R}_i(t) \end{aligned}$$

is not positive. Finally, the estimate $\text{TV}[\rho^n(0, \cdot)] \leq \text{TV}[\bar{\rho}]$ is a simple exercise.

The second way to achieve compactness is via the following *discrete Oleinik-type inequality*. Here, we require (V2) in place of (I2).

Proposition 2 ([30, Proposition 3.2]) *Assume (I1), (V1) and (V2). Then, $(x_i^n)_{i=0}^n$ satisfies for any $t > 0$*

$$\frac{\dot{x}_{i+1}^n(t) - \dot{x}_i^n(t)}{x_{i+1}^n(t) - x_i^n(t)} \leq \frac{1}{t}, \quad i \in \{0, \dots, n-1\}. \quad (15)$$

Proof (15) is equivalent to

$$z_i(t) \doteq t R_i(t) [\dot{x}_{i+1}(t) - \dot{x}_i(t)] \leq \ell_n \quad \text{for all } t > 0, \quad i \in \{1, \dots, n-2\}.$$

We prove the above estimate inductively on i by using (14). Since $z_{n-2}(0) = 0$ and

$$\dot{z}_{n-2} \leq R_{n-2} [v_{\max} - v(R_{n-2})] \left[1 - \frac{z_{n-2}}{\ell_n} \right],$$

a simple comparison argument shows that $z_{n-1}(t)\ell_n$ for all $t \geq 0$, see [30, Proposition 3.2]. Next, we prove that if $z_{i+1}(t) \leq \ell_n$ for all $t \geq 0$ and for some $i \in \{1, \dots, n-2\}$, then $z_i(t) = t R_i(t) [v(R_{i+1}(t)) - v(R_i(t))] \leq \ell_n$ for all $t \geq 0$. Observe that $\text{sign}_+(z_i) = \text{sign}_+(v(R_{i+1}) - v(R_i)) = \text{sign}_+(R_i - R_{i+1})$ for all $i \in \{1, \dots, n-3\}$, where $(z)_+ \doteq \max\{z, 0\}$. The inequality $z_{i+1} \leq \ell_n$ and (V2) implies

$$\frac{d}{dt} (z_i)_+ \leq R_i \left[(v(R_{i+1}) - v(R_i))_+ - v'(R_i) R_i \right] \left(1 - \frac{(z_i)_+}{\ell_n} \right).$$

We observe that the term in the squared bracket in the above estimate is nonnegative. Therefore, again a comparison argument shows that $z_i(t) \leq \ell_n$ for all $t \geq 0$.

Remark 2 We point out that for $i \in \{1, \dots, n-2\}$ the estimate (15) reads

$$\frac{v(R_{i+1}^n(t)) - v(R_i^n(t))}{x_{i+1}^n(t) - x_i^n(t)} \leq \frac{1}{t},$$

which recalls the one-sided Lipschitz condition in [47, 59], which characterizes the entropy solutions to (3).

Remark 3 The result in Proposition 2 implies a uniform bound for $(\rho^n)_n$ in $\mathbf{BV}_{\text{loc}}(\mathbb{R}_+ \times \mathbb{R})$. In this sense, the $\mathbf{L}^\infty \rightarrow \mathbf{BV}$ smoothing effect featured by genuinely nonlinear scalar conservation laws is intrinsically encoded in the particle scheme (11). We omit the details of the proof, and refer to [30, Proposition 3.3].

We prove now (H2'), namely we provide a uniform time continuity estimate in the scaled 1-Wasserstein distance $W_{L,1}$ defined in (8), which ensures strong \mathbf{L}^1 compactness with respect to both space and time.

Proposition 3 *Assume (I1) and (V1). Then the sequence $(\rho^n)_n$ satisfies (H2').*

Proof By (13) and (7), we have that

$$X_{\rho^n(t,\cdot)}(z) = \sum_{i=0}^{n-1} \left[x_i^n(t) + (z - i \ell_n) R_i^n(t)^{-1} \right] \mathbf{1}_{[i \ell_n, (i+1) \ell_n)}(z).$$

For any $0 < s < t$, by (8), (14), and (12)

$$\begin{aligned} W_{L,1}(\rho^n(t, \cdot), \rho^n(s, \cdot)) &\leq v_{\max} |t - s| + \sum_{i=1}^{n-2} \frac{\ell_n}{2} \int_s^t |v(R_{i+1}^n(\tau)) - v(R_i^n(\tau))| d\tau \\ &\quad + \frac{\ell_n}{2} \int_s^t |v_{\max} - v(R_{n-2}^n(\tau))| d\tau \leq v_{\max} |t - s|, \end{aligned}$$

and this concludes the proof.

2.3 Convergence to Entropy Solutions

If besides (I1) and (V1), we assume either (I2) or (V2), then the Propositions 1 and 2 show that $(\rho^n)_n$ satisfies (H1) of Theorem 1 on every time interval $[\delta, T]$ with $0 < \delta < T$. Proposition 3 then implies that $(\rho^n)_n$ satisfies (H2'), hence also (H2) of Theorem 1. Thus, by a simple diagonal argument stretching the time interval $[\delta, T]$ to $(0, T]$, we get that $(\rho^n)_n$ converges (up to a subsequence) a.e. and in $\mathbf{L}_{\text{loc}}^1$ on $(0, T) \times \mathbb{R}$. Let ρ be such limit.

Step 1: ρ is a weak solution to (9). Let $\varphi \in \mathbf{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. By (13), we compute

$$\begin{aligned} &\iint_{\mathbb{R}_+ \times \mathbb{R}} [\rho^n(t, x) \varphi_t(t, x) + f(\rho^n(t, x)) \varphi_x(t, x)] dx dt + \int_{\mathbb{R}} \rho^n(0, x) \varphi(0, x) dx \\ &= \sum_{i=0}^{n-1} \int_{\mathbb{R}_+} \left[-\dot{R}_i^n(t) \left(\int_{x_i^n(t)}^{x_{i+1}^n(t)} \varphi(t, x) dx \right) + R_i^n(t) [\dot{x}_i^n(t) - v(R_i^n(t))] \varphi(t, x_i^n(t)) \right. \\ &\quad \left. - \frac{R_i^n(t)^2}{\ell_n} [\dot{x}_{i+1}^n(t) - v(R_i^n(t))] \left[\int_{x_i^n(t)}^{x_{i+1}^n(t)} \varphi(t, x) dx \right] \right] dt. \end{aligned}$$

Assuming that $\text{supp}[\varphi] \subset [\delta, T] \times \mathbb{R}$ for some $0 < \delta < T$, we obtain

$$\begin{aligned} &\left| \iint_{\mathbb{R}_+ \times \mathbb{R}} [\rho^n(t, x) \varphi_t(t, x) + f(\rho^n(t, x)) \varphi_x(t, x)] dx dt \right| \\ &\leq \frac{T \text{Lip}[\varphi] \ell_n}{2} \left[v_{\max} + \sup_{t \in [\delta, T]} \text{TV}(v(\rho^n(t, \cdot)); J(T)) \right], \end{aligned} \tag{\clubsuit}$$

where $J(T) \doteq [\min\{\text{supp}[\bar{\rho}]\} + v(R)T, \max\{\text{supp}[\bar{\rho}]\} + v_{\max}T]$. Hence, by Proposition 1, the right-hand side in (clubsuit) tends to zero as $n \rightarrow +\infty$, and since ρ^n tends (up to a subsequence) to ρ a.e., we have that ρ is a weak solution to the Cauchy problem (9) for positive times. By (10) and the definition of R_i^n , we have that

$$\begin{aligned} & \left| \int_{\mathbb{R}} [\bar{\rho}(x) - \rho^n(0, x)] \varphi(0, x) dx \right| \\ & \leq 2\ell_n \|\varphi(0, \cdot)\|_{L^\infty(\mathbb{R})} + \sum_{i=0}^{n-1} \left| \int_{\tilde{x}_i^n}^{\tilde{x}_{i+1}^n} \bar{\rho}(x) \left[\varphi(0, x) - \int_{\tilde{x}_i^n}^{\tilde{x}_{i+1}^n} \varphi(0, y) dy \right] dx \right| \end{aligned}$$

and clearly the above quantity goes to zero as $n \rightarrow +\infty$.

Step 2: ρ is an entropy solution to (9). Let $\varphi \in C_c^\infty(\mathbb{R}_+ \times \mathbb{R})$ with $\varphi \geq 0$ and $k \geq 0$. By (13)

$$\begin{aligned} & \iint_{\mathbb{R}_+ \times \mathbb{R}} \left[|\rho(t, x) - k| \varphi_t(t, x) + \text{sign}(\rho(t, x) - k) [f(\rho(t, x)) - f(k)] \varphi_x(t, x) \right] dx dt \\ & = k \int_{\mathbb{R}_+} \left[[v(k) - \dot{x}_0^n(t)] \varphi(t, x_0^n(t)) - [v(k) - \dot{x}_n^n(t)] \varphi(t, x_n^n(t)) \right] dt \\ & \quad + \sum_{i=0}^{n-1} \int_{\mathbb{R}_+} \text{sign}(R_i^n(t) - k) \left[-\dot{R}_i^n(t) \left(\int_{x_i^n(t)}^{x_{i+1}^n(t)} \varphi(t, x) dx \right) \right. \\ & \quad \left. - [R_i^n(t) [\dot{x}_{i+1}^n(t) - v(R_i^n(t))] - k [\dot{x}_{i+1}^n(t) - v(k)]] \varphi(t, x_{i+1}^n(t)) \right. \\ & \quad \left. + [R_i^n(t) [\dot{x}_i^n(t) - v(R_i^n(t))] - k [\dot{x}_i^n(t) - v(k)]] \varphi(t, x_i^n(t)) \right] dt. \end{aligned}$$

Now, we use the equations (14) and (11) as follows. Assuming that $\text{supp}[\varphi] \subset [\delta, T] \times \mathbb{R}$ for some $0 < \delta < T$, we obtain

$$\begin{aligned} & \iint_{\mathbb{R}_+ \times \mathbb{R}} \left[|\rho(t, x) - k| \varphi_t(t, x) + \text{sign}(\rho(t, x) - k) [f(\rho(t, x)) - f(k)] \varphi_x(t, x) \right] dx dt \\ & = k \int_{\mathbb{R}_+} \left[[v(k) - v(R_0^n(t))] \varphi(t, x_0^n(t)) - [v(k) - v_{\max}] \varphi(t, x_n^n(t)) \right] dt \\ & \quad + \sum_{i=0}^{n-2} \int_{\mathbb{R}_+} \text{sign}(R_i^n(t) - k) \\ & \quad \times \left[\frac{R_i^n(t)^2}{\ell_n} [v(R_{i+1}^n(t)) - v(R_i^n(t))] \left[\int_{x_i^n(t)}^{x_{i+1}^n(t)} [\varphi(t, x) - \varphi(t, x_{i+1}^n(t))] dx \right] \right. \\ & \quad \left. + k [[v(R_{i+1}^n(t)) - v(k)] \varphi(t, x_{i+1}^n(t)) - [v(R_i^n(t)) - v(k)] \varphi(t, x_i^n(t))] \right] dt \\ & \quad + \int_{\mathbb{R}_+} \text{sign}(R_{n-1}^n(t) - k) \\ & \quad \times \left[\frac{R_{n-1}^n(t)^2}{\ell_n} [v_{\max} - v(R_{n-1}^n(t))] \left[\int_{x_{n-1}^n(t)}^{x_n^n(t)} [\varphi(t, x) - \varphi(t, x_n^n(t))] dx \right] \right. \\ & \quad \left. + k [[v_{\max} - v(k)] \varphi(t, x_n^n(t)) - [v(R_{n-1}^n(t)) - v(k)] \varphi(t, x_{n-1}^n(t))] \right] dt. \end{aligned}$$

We already proved, see (♣), that

$$\begin{aligned}
& \sum_{i=0}^{n-2} \int_{\mathbb{R}_+} \text{sign}(R_i^n(t) - k) \frac{R_i^n(t)^2}{\ell_n} \left[v(R_{i+1}^n(t)) - v(R_i^n(t)) \right] \\
& \quad \times \left[\int_{x_i^n(t)}^{x_{i+1}^n(t)} [\varphi(t, x) - \varphi(t, x_{i+1}^n(t))] dx \right] dt \\
& + \int_{\mathbb{R}_+} \text{sign}(R_{n-1}^n(t) - k) \frac{R_{n-1}^n(t)^2}{\ell_n} \left[v_{\max} - v(R_{n-1}^n(t)) \right] \\
& \quad \times \left[\int_{x_{n-1}^n(t)}^{x_n^n(t)} [\varphi(t, x) - \varphi(t, x_n^n(t))] dx \right] dt
\end{aligned}$$

converges to zero as $n \rightarrow +\infty$. Hence, to conclude it suffices to observe that

$$\begin{aligned}
& k \left[[v(k) - v(R_0^n(t))] \varphi(t, x_0^n(t)) - [v(k) - v_{\max}] \varphi(t, x_n^n(t)) \right. \\
& + \sum_{i=0}^{n-2} \text{sign}(R_i^n(t) - k) \\
& \quad \times \left[[v(R_{i+1}^n(t)) - v(k)] \varphi(t, x_{i+1}^n(t)) - [v(R_i^n(t)) - v(k)] \varphi(t, x_i^n(t)) \right] \\
& + \text{sign}(R_{n-1}^n(t) - k) \\
& \quad \times \left. \left[[v_{\max} - v(k)] \varphi(t, x_n^n(t)) - [v(R_{n-1}^n(t)) - v(k)] \varphi(t, x_{n-1}^n(t)) \right] \right] \\
& = k \left[\sum_{i=1}^{n-1} [\text{sign}(R_{i-1}^n(t) - k) - \text{sign}(R_i^n(t) - k)] [v(R_i^n(t)) - v(k)] \varphi(t, x_i^n(t)) \right. \\
& \quad + [1 + \text{sign}(R_0^n(t) - k)] [v(k) - v(R_0^n(t))] \varphi(t, x_0^n(t)) \\
& \quad \left. + [1 + \text{sign}(R_{n-1}^n(t) - k)] [v_{\max} - v(k)] \varphi(t, x_n^n(t)) \right] \geq 0.
\end{aligned}$$

3 The LWR Model with Dirichlet Boundary Conditions

In this section, we tackle a new problem in the context of the FTL approximation for traffic flow models, namely the approximation of the IBVP with time-varying Dirichlet boundary conditions

$$\begin{cases} \rho_t + [\rho v(\rho)]_x = 0, & (t, x) \in \mathbb{R}_+ \times \mathcal{Q}, \\ \rho(0, x) = \bar{\rho}(x), & x \in \mathcal{Q}, \\ \rho(t, 0) = \bar{\rho}_0(t), & t \in \mathbb{R}_+, \\ \rho(t, 1) = \bar{\rho}_1(t), & t \in \mathbb{R}_+, \end{cases} \quad (16)$$

where, for notational simplicity, we let $\mathcal{Q} \doteq (0, 1)$. We assume that the velocity map satisfies (V1); further, we assume that there exists $\delta > 0$ such that the initial datum and the boundary data satisfy, respectively,

$$\bar{\rho} \in \mathbf{L}^\infty \cap \mathbf{BV}(\mathcal{Q}; [\delta, \rho_{\max}]), \quad (\text{I3})$$

$$\bar{\rho}_0, \bar{\rho}_1 \in \mathbf{L}^\infty \cap \mathbf{Lip} \cap \mathbf{BV}(\mathbb{R}_+; [\delta, \rho_{\max}]). \quad (\text{B})$$

We adapt the definition of entropy solution given in [25, Definition 2.1], see also [1, 36], to the case under consideration.

Definition 2 Assume (I3), (B) and (V1). We say that $\rho \in \mathbf{C}^0(\mathbb{R}_+; \mathbf{L}_{\text{loc}}^\infty(\bar{\mathcal{Q}}; [0, \rho_{\max}]))$ is an *entropy solution* to the IBVP (16) if

- for any test function $\phi \in \mathbf{C}_c^\infty(\mathbb{R}_+ \times \mathcal{Q})$ with $\phi \geq 0$ and for any $k \in [0, \rho_{\max}]$

$$0 \leq \iint_{\mathbb{R}_+ \times \mathcal{Q}} \left[|\rho - k| \phi_t + \text{sign}(\rho - k) [f(\rho) - f(k)] \phi_x \right] dx dt \\ + \int_{\mathcal{Q}} |\bar{\rho} - k| \phi(0, x) dx;$$

- for a.e. $\tau \geq 0$, we have $\rho(\tau, 0^+) = u(1, x)$ for all $x > 0$, where u is the self-similar Lax solution to the Riemann problem

$$\begin{cases} u_t + f(u)_x = 0, & (t, x) \in \mathbb{R}_+ \times \mathbb{R}, \\ u(0, x) = \begin{cases} \bar{\rho}_0(\tau) & \text{if } x < 0, \\ \rho(\tau, 0^+) & \text{if } x > 0, \end{cases} & x \in \mathbb{R}; \end{cases}$$

- for a.e. $\tau \geq 0$, we have $\rho(\tau, 1^-) = w(1, x)$ for all $x < 0$, where w is the self-similar Lax solution to the Riemann problem

$$\begin{cases} w_t + f(w)_x = 0, & (t, x) \in \mathbb{R}_+ \times \mathbb{R}, \\ w(0, x) = \begin{cases} \rho(\tau, 1^-) & \text{if } x < 0, \\ \bar{\rho}_1(\tau) & \text{if } x > 0, \end{cases} & x \in \mathbb{R}. \end{cases}$$

3.1 The Follow-the-Leader Scheme and Main Result

We now introduce rigorously our FTL approximation scheme for (16). Assume (I3), (B) and (V1). For a given $T > 0$ and an integer $m \in \mathbb{N}$, we set $\tau_m \doteq T/m$. We approximate the boundary data $\bar{\rho}_0, \bar{\rho}_1$ with $\bar{\rho}_{0,m}, \bar{\rho}_{1,m}$ defined by

$$\bar{\rho}_{i,m} \doteq \sum_{k=0}^{m-1} \bar{\rho}_i^k \mathbf{1}_{[k\tau_m, (k+1)\tau_m]} \quad \text{with } \bar{\rho}_i^k \doteq \bar{\rho}_i(k\tau_m), \quad i \in \{0, 1\}.$$

Let again $L \doteq \|\bar{\rho}\|_{L^1(\Omega)}$ and $R \doteq \|\bar{\rho}\|_{L^\infty(\Omega)}$. Fix $n \in \mathbb{N}$ sufficiently large, and set $\ell_n \doteq L/n$. Let $\bar{x}_0^0, \dots, \bar{x}_n^0$ be defined recursively by

$$\begin{cases} \bar{x}_0^0 \doteq 0, \\ \bar{x}_i^0 \doteq \sup \left\{ x \in \Omega : \int_{\bar{x}_{i-1}^0}^x \bar{\rho}_\delta(x) dx < \ell_n \right\}, \quad i \in \{1, \dots, n\}. \end{cases}$$

By construction $\bar{x}_n^0 = 1$. We introduce the *artificial queuing mass* Q and the *number of queuing particles* N defined by

$$Q \doteq 2T v_{\max} \rho_{\max}, \quad N \doteq \lceil Q/\ell_n \rceil. \quad (17)$$

Let the initial positions of the *queuing particles* $\bar{x}_{-N}^0, \dots, \bar{x}_{-1}^0$ be defined by

$$\begin{cases} \bar{x}_i^0 \doteq i \frac{\ell_n}{\bar{\rho}_0^0}, & i \in \{-N+1, \dots, -1\}, \\ \bar{x}_{-N}^0 \doteq \bar{x}_{-N+1}^0 - \frac{q_n}{\bar{\rho}_0^0}, \end{cases}$$

where $q_n \doteq Q - \ell_n(N-1) \in [0, \ell_n]$ and $\bar{\rho}_0^0 \doteq \bar{\rho}_0(0)$. The queuing particles are set in \mathbb{R}_- , with equal distances from each other in order to match the density $\bar{\rho}_0^0$, with the only exception of the leftmost particle \bar{x}_{-N}^0 , which carries a mass q_n (possibly less than ℓ_n) in order to have a fixed total mass Q for the whole set of queuing particles.

The $(n+N+1)$ positions $\bar{x}_{-N}^0, \dots, \bar{x}_n^0$ are taken as initial conditions of the FTL system

$$\begin{cases} \dot{x}_i^0(t) = v(R_i^0(t)), & t \in [0, \tau_m], \quad i \in \{-N, \dots, n-1\}, \\ \dot{x}_n^0(t) = v(\bar{\rho}_1^0), & t \in [0, \tau_m], \\ x_i^0(0) = \bar{x}_i^0, & i \in \{-N, \dots, n\}, \end{cases}$$

where we have denoted

$$\begin{cases} R_i^0(t) \doteq \frac{\ell_n}{x_{i+1}(t) - x_i(t)}, & t \in [0, \tau_m], \quad i \in \{-N+1, \dots, n-1\}, \\ R_{-N}^0(t) \doteq \frac{q_n}{x_{-N+1}(t) - x_{-N}(t)}, & t \in [0, \tau_m]. \end{cases}$$

We then extend the above definitions to $[0, T]$ recursively as follows. For any $k \in \mathbb{N}$ with $k \geq 1$, we denote by h_0^k the number of particles that *strictly* crossed $x = 0$ during the time interval $(0, k\tau_m]$, and by h_1^k the number of particles that crossed $x = 1$ during the same time interval (counting the possible particle positioned at $x = 1$ at time $k\tau_m$). We rearrange the particles positions at time $t = k\tau_m$ by setting

$$\bar{x}_i^k \doteq \begin{cases} x_i(k \tau_m), & i \in \{-h_0^k - 1, \dots, n - h_1^k + 1\}, \\ x_{n-h_1^k+1}(k \tau_m) + (i - n + h_1^k - 1) \frac{\ell_n}{\bar{\rho}_1^k}, & i \in \{n - h_1^k + 2, \dots, n\}, \\ x_{-h_0^k-1}(k \tau_m) + (i + h_0^k + 1) \frac{\ell_n}{\bar{\rho}_0^k}, & i \in \{-N + 1, \dots, -h_0^k - 2\}, \\ \bar{x}_{-N+1}^k - \frac{q_n}{\bar{\rho}_0^k}, & i = -N. \end{cases}$$

In other words, we maintain the same position for all the particles that are positioned in Ω , plus the rightmost particle in $(-\infty, 0]$ and the leftmost particle in $[1, +\infty)$, and we move all the other particles to make them equidistant in order to match the updated boundary conditions for the density $\bar{\rho}_0^k$ and $\bar{\rho}_1^k$.

Then, the $(n + N + 1)$ positions $\bar{x}_{-N}^k, \dots, \bar{x}_n^k$ are taken as initial conditions of the following FTL system

$$\begin{cases} \dot{x}_i^k(t) = v(R_i^k(t)), & t \in [k \tau_m, (k + 1) \tau_m], i \in \{-N, \dots, n - 1\}, \\ \dot{x}_n^k(t) = v(\bar{\rho}_1^k), & t \in [k \tau_m, (k + 1) \tau_m], \\ x_i^k(k \tau_m) = \bar{x}_i^k, & i \in \{-N, \dots, n\}, \end{cases} \quad (18)$$

where we have denoted

$$\begin{cases} R_i^k(t) \doteq \frac{\ell_n}{x_{i+1}(t) - x_i(t)}, & t \in [k \tau_m, (k + 1) \tau_m], i \in \{-N + 1, \dots, n - 1\}, \\ R_{-N}^k(t) \doteq \frac{q_n}{x_{-N+1}(t) - x_{-N}(t)}, & t \in [k \tau_m, (k + 1) \tau_m]. \end{cases}$$

We observe that the number of queuing particles N has been chosen in order to guarantee that a number of particles of the order $N/2$ (as $n \rightarrow +\infty$) will not cross $x = 0$ within the time interval $[0, T]$.

Remark 4 Our choice for the above particle scheme is motivated as follows. In order to approach the entropy solution according to Definition 2 in the $n \rightarrow +\infty$ limit, on each time interval $[k \tau_m, (k + 1) \tau_m]$ we tend to the entropy solution with constant boundary conditions by ‘extending’ the discrete particle density at time $t = k \tau_m$ in a way to match said boundary conditions (see a similar construction in, e.g., [25]).

The following discrete maximum–minimum principle ensures that the particles strictly preserve their initial order.

Lemma 2 (Discrete maximum–minimum principle) *Assume (I3), (V1) and (B). Then, the solution to (18) satisfies for any $t \geq 0$*

$$\frac{\ell_n}{R} \leq x_{i+1}(t) - x_i(t) \leq \frac{\ell_n}{\delta}, \quad i \in \{-N, \dots, n - 1\}. \quad (19)$$

Proof The lower bound on the time interval $[0, \tau_m]$ is a consequence of the result in Lemma 1. We now prove the upper bound on $[0, \tau_m]$. We consider first $i = n - 1$. By contradiction, assume that there exist $t_1, t_2 \in (0, \tau_m)$ such that $t_1 < t_2$, $x_n(t) -$

$x_{n-1}(t) \leq \ell_n/\delta$ for $t < t_1$, $x_n(t_1) - x_{n-1}(t_1) = \ell_n/\delta$ and $x_n(t) - x_{n-1}(t) > \ell_n/\delta$ for $t \in (t_1, t_2)$. Then, for any $t \in (t_1, t_2)$, we have $\bar{\rho}_1^0 \geq \delta > R_{n-1}^0(t)$ and therefore

$$\begin{aligned} x_n(t) - x_{n-1}(t) &= x_n(t_1) - x_{n-1}(t_1) + \int_{t_1}^t [v(\bar{\rho}_1^0) - v(R_{n-1}^0(s))] \, ds \\ &= \frac{\ell_n}{\delta} + \int_{t_1}^t [v(\bar{\rho}_1^0) - v(R_{n-1}^0(s))] \, ds \leq \frac{\ell_n}{\delta}, \end{aligned}$$

which gives a contradiction. We prove now the upper bound on $[0, \tau_m]$ for all the other vehicles inductively. Assume

$$\sup_{t \in [0, \tau_m]} [x_{i+2}(t) - x_{i+1}(t)] \leq \frac{\ell_n}{\delta},$$

and by contradiction that there exist $t_1, t_2 \in (0, \tau_m)$ such that $t_1 < t_2$, $x_{i+1}(t) - x_i(t) \leq \ell_n/\delta$ for $t < t_1$, $x_{i+1}(t_1) - x_i(t_1) = \ell_n/\delta$ and $x_{i+1}(t) - x_i(t) > \ell_n/\delta$ for $t \in (t_1, t_2)$. Then, for any $t \in (t_1, t_2)$, we have $R_{i+1}^0(t) \geq \delta > R_i^0(t)$ and therefore

$$x_{i+1}(t) - x_i(t) = x_{i+1}(t_1) - x_i(t_1) + \int_{t_1}^t [v(R_{i+1}^0(s)) - v(R_i^0(s))] \, ds \leq \frac{\ell_n}{\delta}$$

which gives a contradiction. This proves the assertion on $[0, \tau_m]$. Now, at each time step $t = k \tau_m$ the set of particles is rearranged outside Ω in such a way that two consecutive particles satisfy

$$\begin{cases} x_{i+1}(k \tau_m) - x_i(k \tau_m) = \frac{\ell_n}{\bar{\rho}_1^k} \leq \frac{\ell_n}{\delta}, & i \in \{n - h_1^k + 1, \dots, n - 1\}, \\ x_{i+1}(k \tau_m) - x_i(k \tau_m) = \frac{\ell_n}{\bar{\rho}_0^k} \leq \frac{\ell_n}{\delta}, & i \in \{-N + 1, \dots, -h_0^k - 1\}, \\ x_{-N+1}(k \tau_m) - x_{-N}(k \tau_m) = \frac{q_n}{\bar{\rho}_0^k} \leq \frac{\ell_n}{\delta}. \end{cases}$$

Inside the domain Ω , the inequalities in (19) are satisfied due to the maximum-minimum principle holding on the previous time interval. Hence, we can reapply inductively the above procedure and easily get the assertion.

We define the discrete density for $t \in (0, T]$ as

$$\rho^{n,m}(t, x) \doteq \sum_{m=0}^{m-1} \sum_{i=-N}^{n-1} R_i^k(t) \mathbf{1}_{[x_i(t), x_{i+1}(t))}(x) \mathbf{1}_{(k \tau_m, (k+1) \tau_m]}(t).$$

It is easy to verify that $\|\rho^{n,m}(t, \cdot)\|_{\mathbf{L}^1(\Omega)} = Q + L$ for all $t \geq 0$. We state the main result of this section, as well as the main novel result of this chapter.

Theorem 4 *Assume (I3), (V1) and (B). Then, $(\rho^{n,m} \mathbf{1}_\Omega)_{n,m}$ converges (up to a subsequence) a.e. and in \mathbf{L}^1 on $\mathbb{R}_+ \times \Omega$ to a weak solution ρ to the IVP (16) in the interior of Ω .*

Our conjecture is that the limit ρ is in fact the unique entropy solution to the IBVP (16) in the sense of Definition 2. This is motivated by the construction of our FTL approximation scheme, which relies on the Definition 2, see Remark 4. Moreover, the numerical simulations performed in Subsection 6.2 suggest it. We rigorously prove the consistency of the scheme only in simple cases in Subsection 3.3 below.

The fact that the limit ρ in the statement of Theorem 4 is a weak solution to the LWR equation in the interior of Ω can be easily proven as in the proof of Theorem 3, and therefore, we omit the details. Hence, we only need to prove convergence of the sequence $\rho^{n,m}$ strongly in \mathbf{L}^1 up to a subsequence. This task is the goal of the next section.

Remark 5 As already explained in the introduction, we recall that the boundary condition does not need to be updated in time as long as no waves coming from Ω hit the boundary $\partial\Omega$. In particular, if $\bar{\rho}$, $\bar{\rho}_0$, and $\bar{\rho}_1$ are constant, the solution is simply obtained as the restriction to Ω of the entropy solution to the Cauchy problem (9) with initial condition $\bar{\rho}_0 \mathbf{1}_{(-\infty,0)} + \bar{\rho} \mathbf{1}_\Omega + \bar{\rho}_1 \mathbf{1}_{(1,+\infty)}$ and no update in time of the boundary data is needed. There are other cases in which $\bar{\rho}$, $\bar{\rho}_0$, and $\bar{\rho}_1$ are not necessarily constant and such situation occurs. We highlight one of them here in the special case $v(\rho) \doteq 1 - \rho$, which yields $f(\rho) \doteq \rho(1 - \rho)$. Indeed, a very simple argument based on the WFT approximation (see, e.g., [25]) shows that for any $h \in [0, 1]$, if we denote with ρ_k the restriction to Ω of the solution to the Riemann problem with initial datum $h \mathbf{1}_{(-\infty,1)} + k \mathbf{1}_{(1,+\infty)}$, then $\rho_k = \rho_{1/2}$ for all $k \in [0, 1/2]$. Arguing in a similar way for $x = 0$, it is easy to see that if the boundary data $\bar{\rho}_0$ and $\bar{\rho}_1$ take values in $[1/2, 1]$ and $[0, 1/2]$, respectively, then no updates of the boundary data is needed.

3.2 Estimates

Similar to Section 2.2, the proof of Theorem 4 is based on some estimates which infer suitable space–time compactness. We now prove the following **BV** estimate for $(\rho^{n,m} \mathbf{1}_\Omega)_{n,m}$.

Proposition 4 *Assume (I3), (V1) and (B). Then for any $t > 0$*

$$\text{TV}(\rho^{n,m}(t, \cdot)) \leq C,$$

where $C \doteq \text{TV}(\bar{\rho}) + \text{TV}(\bar{\rho}_0) + \text{TV}(\bar{\rho}_1) + |\bar{\rho}_0(0^+) - \bar{\rho}(0^+)| + |\bar{\rho}_1(0^+) - \bar{\rho}(1^-)|$.

Proof We define $\Upsilon : (0, T] \rightarrow \mathbb{R}_+$ by letting for any $t \in (k \tau_m, (k+1) \tau_m]$ with $k \in \mathbb{N}$

$$\begin{aligned} \Upsilon(t) &\doteq |R_{-h_0^{k+1}-2}^k(t) - \bar{\rho}_0^k| + |R_{n-h_1^k+1}^k(t) - \bar{\rho}_1^k| + \sum_{i=-h_0^{k+1}-2}^{n-h_1^k} |R_{i+1}^k(t) - R_i^k(t)| \\ &\quad + \sum_{j=k}^{m-1} \left[|\bar{\rho}_0^{j+1} - \bar{\rho}_0^j| + |\bar{\rho}_1^{j+1} - \bar{\rho}_1^j| \right]. \end{aligned}$$

We observe that

$$R_{n-h_1^k+1}^k(t) = R_{n-h_1^k+2}^k(t) = \dots = R_{n-1}^k(t) = \bar{\rho}_1^k \quad \text{for all } t \in [k \tau_m, (k+1) \tau_m],$$

due to the fact that the above quantities are all equal at time $t = k \tau_m$ and the leader x_n travels with speed $v(\bar{\rho}_1^k)$ during the whole time interval.

We claim that $\mathcal{A}\Upsilon(t) \doteq \Upsilon(t^+) - \Upsilon(t^-) \leq 0$ for all $t \in (0, T)$. Let us first consider $t \in (k \tau_m, (k+1) \tau_m)$. In this case,

$$\begin{aligned} \dot{\Upsilon}(t) &= \text{sign}(R_{-h_0^{k+1}-2}^k(t) - \bar{\rho}_0^k) \dot{R}_{-h_0^{k+1}-2}^k(t) + \text{sign}(R_{n-h_1^k+1}^k(t) - \bar{\rho}_1^k) \dot{R}_{n-h_1^k+1}^k(t) \\ &\quad + \sum_{i=-h_0^{k+1}-2}^{n-h_1^k} \text{sign}(R_{i+1}^k(t) - R_i^k(t)) (\dot{R}_{i+1}^k(t) - \dot{R}_i^k(t)) \\ &= \left[\text{sign}(R_{-h_0^{k+1}-2}^k(t) - \bar{\rho}_0^k) - \text{sign}(R_{-h_0^{k+1}-1}^k(t) - R_{-h_0^{k+1}-2}^k(t)) \right] \dot{R}_{-h_0^{k+1}-2}^k(t) \\ &\quad + \left[\text{sign}(R_{n-h_1^k+1}^k(t) - \bar{\rho}_1^k) + \text{sign}(R_{n-h_1^k+1}^k(t) - R_{n-h_1^k}^k(t)) \right] \dot{R}_{n-h_1^k+1}^k(t) \\ &\quad + \sum_{i=-h_0^{k+1}-1}^{n-h_1^k} \left[\text{sign}(R_i^k(t) - R_{i-1}^k(t)) - \text{sign}(R_{i+1}^k(t) - R_i^k(t)) \right] \dot{R}_i^k(t) \leq 0. \end{aligned}$$

The above estimate holds because the quantities

$$\begin{aligned} &\left[\text{sign}(R_{-h_0^{k+1}-2}^k(t) - \bar{\rho}_0^k) - \text{sign}(R_{-h_0^{k+1}-1}^k(t) - R_{-h_0^{k+1}-2}^k(t)) \right] \dot{R}_{-h_0^{k+1}-2}^k(t) \\ &= \left[\text{sign}(\bar{\rho}_0^k - R_{-h_0^{k+1}-2}^k(t)) + \text{sign}(R_{-h_0^{k+1}-1}^k(t) - R_{-h_0^{k+1}-2}^k(t)) \right] \\ &\quad \times \frac{R_{-h_0^{k+1}-2}^k(t)^2}{\ell_n} \left[v(R_{-h_0^{k+1}-1}^k(t)) - v(R_{-h_0^{k+1}-2}^k(t)) \right], \\ &\left[\text{sign}(R_{n-h_1^k+1}^k(t) - \bar{\rho}_1^k) + \text{sign}(R_{n-h_1^k+1}^k(t) - R_{n-h_1^k}^k(t)) \right] \dot{R}_{n-h_1^k+1}^k(t) \\ &= \left[\text{sign}(\bar{\rho}_1^k - R_{n-h_1^k+1}^k(t)) + \text{sign}(R_{n-h_1^k}^k(t) - R_{n-h_1^k+1}^k(t)) \right] \frac{R_{n-h_1^k+1}^k(t)^2}{\ell_n} \\ &\quad \times \left[v(\bar{\rho}_1^k) - v(R_{n-h_1^k+1}^k(t)) \right], \\ &\left[\text{sign}(R_i^k(t) - R_{i-1}^k(t)) - \text{sign}(R_{i+1}^k(t) - R_i^k(t)) \right] \dot{R}_i^k(t) \end{aligned}$$

$$= \left[\text{sign}(R_{i-1}^k(t) - R_i^k(t)) + \text{sign}(R_{i+1}^k(t) - R_i^k(t)) \right] \frac{R_i^k(t)^2}{\ell_n} \left[v(R_{i+1}^k(t)) - v(R_i^k(t)) \right]$$

are not positive. Consider now $t = (k+1) \tau_m$ with $k \in \mathbb{N}$. In this case, using that

$$\begin{aligned} R_{-h_{k+2}-2}^{k+1}(t^+) &= R_{-h_{k+2}-1}^{k+1}(t^+) = \dots = R_{-h_0^{k+1}-2}^{k+1}(t^+) = \bar{\rho}_0^{k+1}, \\ R_{n-h_1^{k+1}+1}^{k+1}(t^+) &= R_{n-h_1^{k+1}+2}^{k+1}(t^+) = \dots = R_{n-h_1^k+1}^{k+1}(t^+) = \bar{\rho}_1^{k+1}, \end{aligned}$$

we easily obtain

$$\begin{aligned} \Delta \Upsilon(t) &= -|R_{-h_0^{k+1}-2}^k(t) - \bar{\rho}_0^k| - |R_{n-h_1^k+1}^k(t) - \bar{\rho}_1^k| \\ &\quad + |R_{n-h_1^{k+1}+1}^{k+1}(t) - R_{n-h_1^{k+1}}(t)| - |R_{n-h_1^{k+1}+1}^k(t) - R_{n-h_1^{k+1}}(t)| \\ &\quad - \sum_{j=1}^{h_1^{k+1}-h_1^k} |R_{n-h_1^{k+1}+j+1}^k(t) - R_{n-h_1^{k+1}+j}^k(t)| \\ &\quad + |R_{-h_0^{k+1}-1}^k(t) - R_{-h_0^{k+1}-2}^{k+1}(t)| - |R_{-h_0^{k+1}-1}^k(t) - R_{-h_0^{k+1}-2}^k(t)| \\ &\quad - |\bar{\rho}_0^{k+1} - \bar{\rho}_0^k| - |\bar{\rho}_1^{k+1} - \bar{\rho}_1^k| \\ &= \left[|R_{-h_0^{k+1}-1}^k(t) - \bar{\rho}_0^{k+1}| - |R_{-h_0^{k+1}-1}^k(t) - R_{-h_0^{k+1}-2}^k(t)| - |R_{-h_0^{k+1}-2}^k(t) - \bar{\rho}_0^k| \right. \\ &\quad \left. - |\bar{\rho}_0^k - \bar{\rho}_0^{k+1}| \right] + \left[|R_{n-h_1^k+1}^k(t) - \bar{\rho}_1^{k+1}| - |\bar{\rho}_1^k - \bar{\rho}_1^{k+1}| - |R_{n-h_1^k+1}^k(t) - \bar{\rho}_1^k| \right. \\ &\quad \left. - \sum_{j=0}^{h_1^{k+1}-h_1^k} |R_{n-h_1^{k+1}+j+1}^k(t) - R_{n-h_1^{k+1}+j}^k(t)| \right] \leq 0, \end{aligned}$$

where the last inequality follows by a simple triangular inequality. In conclusion, we have that

$$\begin{aligned} \text{TV}(\rho^{n,m}(t, \cdot)) &\leq \Upsilon(t) \leq \Upsilon(0^+) = \left| \frac{\ell_n}{1 - \bar{x}_{n-1}^0} - \bar{\rho}_1^0 \right| \\ &\quad + \sum_{i=-1}^{n-2} \left| \frac{\ell_n}{\bar{x}_{i+2}^0 - \bar{x}_{i+1}^0} - \frac{\ell_n}{\bar{x}_{i+1}^0 - \bar{x}_i^0} \right| + \sum_{j=0}^{m-1} \left[|\bar{\rho}_0^{j+1} - \bar{\rho}_0^j| + |\bar{\rho}_1^{j+1} - \bar{\rho}_1^j| \right] \\ &\leq \left| \int_{\bar{x}_{n-1}^0}^1 \bar{\rho}(x) \, dx - \bar{\rho}_1(0) \right| + \sum_{i=0}^{n-2} \left| \int_{\bar{x}_{i+1}^0}^{\bar{x}_{i+2}^0} \bar{\rho}(x) \, dx - \int_{\bar{x}_i^0}^{\bar{x}_{i+1}^0} \bar{\rho}(x) \, dx \right| \\ &\quad + \left| \int_0^{\bar{x}_1^0} \bar{\rho}(x) \, dx - \bar{\rho}_0(0) \right| + \text{TV}(\bar{\rho}_0) + \text{TV}(\bar{\rho}_1) \leq C. \end{aligned}$$

We provide now a uniform time continuity estimate with respect to the rescaled 1-Wasserstein distance $W_{Q+L,1}$ defined in (8).

Proposition 5 Assume (I3), (V1) and (B). Then the sequence $(\rho^{n,m})_{n,m}$ satisfies (H2), which in the present framework writes

$$\lim_{h \downarrow 0} \left[\sup_{n,m \in \mathbb{N}} \left[\int_0^{T-h} W_{Q+L,1}(\rho^{n,m}(t+h), \rho^{n,m}(t)) dt \right] \right] = 0. \quad (20)$$

Proof For simplicity, we drop the indexes n, m in the notation and use W_1 instead of $W_{Q+L,1}$. The above Wasserstein distance is computed via the pseudo-inverse variable

$$\begin{aligned} X_{\rho(t,\cdot)}(z) &= \left[x_{-N}(t) + z R_{-N}(t)^{-1} \right] \mathbf{1}_{[0,q_n]}(z) \\ &+ \sum_{i=-N+1}^{n-1} \left[x_i(t) + \left(z - (q_n + (i+N-1)\ell_n) \right) R_i(t)^{-1} \right] \mathbf{1}_{[q_n + (i+N-1)\ell_n, q_n + (i+N)\ell_n]}(z). \end{aligned}$$

We recall that for all $t \geq 0$, $X_{\rho(t,\cdot)}$ is a strictly increasing function on $[0, Q+L]$.

For $k \tau_m < s < t < (k+1) \tau_m$, we compute

$$\begin{aligned} W_1(\rho(t,\cdot), \rho(s,\cdot)) &= \|X_{\rho(t,\cdot)} - X_{\rho(s,\cdot)}\|_{\mathbf{L}^1([0,Q+L])} \\ &= \int_0^{q_n} \left| x_{-N}(t) - x_{-N}(s) + z (R_{-N}(t)^{-1} - R_{-N}(s)^{-1}) \right| dz \\ &\quad + \sum_{i=-N+1}^{n-1} \int_0^{\ell_n} \left| x_i(t) - x_i(s) + z (R_i(t)^{-1} - R_i(s)^{-1}) \right| dz \\ &\leq q_n |x_{-N}(t) - x_{-N}(s)| + |R_{-N}(t)^{-1} - R_{-N}(s)^{-1}| \int_0^{q_n} z dz \\ &\quad + \sum_{i=-N+1}^{n-1} \ell_n |x_i(t) - x_i(s)| + \sum_{i=-N+1}^{n-1} \left| R_i(t)^{-1} - R_i(s)^{-1} \right| \int_0^{\ell_n} z dz \\ &\leq (Q+L) v_{\max} |t-s| + \frac{q_n^2}{2} \int_s^t \left| \frac{d}{d\tau} \left[\frac{1}{R_{-N}(\tau)} \right] \right| d\tau + \sum_{i=-N+1}^{n-1} \frac{\ell_n^2}{2} \int_s^t \left| \frac{d}{d\tau} \left[\frac{1}{R_i(\tau)} \right] \right| d\tau \\ &= (Q+L) v_{\max} |t-s| + \frac{q_n}{2} \int_s^t |v(R_{-N+1}(\tau)) - v(R_{-N}(\tau))| d\tau \\ &\quad + \sum_{i=-N}^{n-2} \frac{\ell_n}{2} \int_s^t |v(R_{i+1}(\tau)) - v(R_i(\tau))| d\tau + \frac{\ell_n}{2} \int_s^t |v(\bar{\rho}_1^k) - v(R_{n-2}(\tau))| d\tau \\ &\leq \frac{3}{2} (Q+L) v_{\max} (t-s). \end{aligned} \quad (21)$$

As a consequence of the above computation, the curve $[0, T] \ni t \mapsto \rho^{n,m}(t, \cdot)$ is equicontinuous in the W_1 -topology on open intervals of the form $(k \tau_m, (k+1) \tau_m)$ and

$$W_1(\rho(((k+1) \tau_m)^-, \cdot), \rho((k \tau_m)^+, \cdot)) \leq C \tau_m, \quad k \in \mathbb{N}, \quad (22)$$

where C is some positive constant independent of n, m , and h . On the other hand, due to the rearrangements of the particles outside \mathcal{Q} at each time step, such curve may feature a jump discontinuity. Let $t = (k+1)\tau_m$. We estimate the jump

$$\begin{aligned} W_1(\rho(t^+, \cdot), \rho(t^-, \cdot)) &= \|X_{\rho(t^+, \cdot)} - X_{\rho(t^-, \cdot)}\|_{\mathbf{L}^1([0, L])} \leq q_n |x_{-N}(t^+) - x_{-N}(t^-)| \\ &+ \sum_{i=-N+1}^{-h_0^{k+1}-2} \ell_n |x_i(t^+) - x_i(t^-)| + \sum_{i=n-h_1^{k+1}+2}^{n-1} \ell_n |x_i(t^+) - x_i(t^-)| \\ &+ |(\bar{\rho}_0^{k+1})^{-1} - R_{-N}(t^-)^{-1}| \frac{\ell_n^2}{2} + \sum_{i=-N+1}^{-h_0^{k+1}-2} |(\bar{\rho}_0^{k+1})^{-1} - R_i(t^-)^{-1}| \frac{\ell_n^2}{2} \\ &+ \sum_{i=n-h_1^{k+1}+1}^{n-1} |(\bar{\rho}_1^{k+1})^{-1} - R_i(t^-)^{-1}| \frac{\ell_n^2}{2}, \end{aligned} \quad (23)$$

where we have used the fact that $t \mapsto R_i(t)^{-1}$ is continuous for all $i \in \{-h_0^{k+1} - 1, \dots, n - h_1^{k+1}\}$. We claim that for any $i \in \{-N, \dots, -h_0^{k+1} - 2\}$, we have the estimate

$$|x_i(t^+) - x_i(t^-)| \leq \left[2v_{\max} + \frac{Q}{\delta^2} \text{Lip}(\bar{\rho}_0) \right] \tau_m. \quad (24)$$

Indeed, for any $i \in \{-N + 1, \dots, -h_0^{k+1} - 2\}$, we have

$$\begin{aligned} &|x_i(t^+) - x_i(t^-)| \leq |x_i(t^+) - x_i((t - \tau_m)^+)| + |x_i((t - \tau_m)^+) - x_i(t^-)| \\ &= \left| x_{-h_0^{k+1}-1}(t) + (i + h_0^{k+1} + 1) \frac{\ell_n}{\bar{\rho}_0^{k+1}} - x_{-h_0^{k+1}-1}((t - \tau_m)^+) - (i + h_0^{k+1} + 1) \frac{\ell_n}{\bar{\rho}_0^k} \right| \\ &\quad + \int_{t-\tau_m}^t v(R_i(s)) \, ds \\ &\leq \int_{t-\tau_m}^t v(R_{-h_0^{k+1}-1}(s)) \, ds + \ell_n |i + h_0^{k+1} + 1| \left| \frac{1}{\bar{\rho}_0^{k+1}} - \frac{1}{\bar{\rho}_0^k} \right| + \int_{t-\tau_m}^t v(R_i(s)) \, ds \\ &\leq 2\tau_m v_{\max} + Q \left| \frac{1}{\bar{\rho}_0^{k+1}} - \frac{1}{\bar{\rho}_0^k} \right| \leq \left[2v_{\max} + \frac{Q}{\delta^2} \text{Lip}(\bar{\rho}_0) \right] \tau_m, \end{aligned}$$

and analogously

$$\begin{aligned} &|x_{-N}(t^+) - x_{-N}(t^-)| \leq |x_{-N}(t^+) - x_{-N}((k\tau_m)^+)| + |x_{-N}((k\tau_m)^+) - x_{-N}(t^-)| \\ &= \left| x_{-h_0^{k+1}-1}(t) + (i + h_0^{k+1} + 1) \frac{\ell_n}{\bar{\rho}_0^{k+1}} - \frac{q_n}{\bar{\rho}_0^{k+1}} \right. \\ &\quad \left. - x_{-h_0^{k+1}-1}((t - \tau_m)^+) - (i + h_0^{k+1} + 1) \frac{\ell_n}{\bar{\rho}_0^k} + \frac{q_n}{\bar{\rho}_0^k} \right| + \int_{t-\tau_m}^t v(R_i(s)) \, ds \end{aligned}$$

$$\begin{aligned} &\leq \int_{t-\tau_m}^t \left[v(R_{-h_0^{k+1}-1}(s)) + v(R_i(s)) \right] ds + [\ell_n |i + h_0^{k+1} + 1| + q_n] \left| \frac{1}{\bar{\rho}_0^{k+1}} - \frac{1}{\bar{\rho}_0^k} \right| \\ &\leq 2 \tau_m v_{\max} + Q \left| \frac{1}{\bar{\rho}_0^{k+1}} - \frac{1}{\bar{\rho}_0^k} \right| \leq \left[2 v_{\max} + \frac{Q}{\delta^2} \text{Lip}(\bar{\rho}_0) \right] \tau_m. \end{aligned}$$

Moreover, for all $i \in \{-N+1, \dots, -h_0^{k+1}-2\}$,

$$\begin{aligned} |(\bar{\rho}_0^{k+1})^{-1} - R_i(t^-)^{-1}| &\leq |(\bar{\rho}_0^{k+1})^{-1} - (\bar{\rho}_0^k)^{-1}| + |(R_i((t - k \tau_m)^+)^{-1} - R_i(t^-)^{-1}| \\ &\leq |(\bar{\rho}_0^{k+1})^{-1} - (\bar{\rho}_0^k)^{-1}| + \frac{1}{\ell_n} \int_{t-\tau_m}^t |v(R_{i+1}(s)) - v(R_i(s))| ds \\ &\leq \left[\frac{v_{\max}}{\ell_n} + \frac{1}{\delta^2} \text{Lip}(\bar{\rho}_0) \right] \tau_m, \end{aligned} \quad (25)$$

and the same estimate holds for $i = -N$ with q_n replacing ℓ_n . For any $i \in \{n - h_1^k + 2, \dots, n-1\}$, we estimate

$$\begin{aligned} &|x_i(t^+) - x_i(t^-)| \leq |x_i(t^+) - x_i((t - \tau_m)^+)| + |x_i((t - \tau_m)^+) - x_i(t^-)| \\ &= \left| x_{n-h_1^{k+1}+1}(t) + (i - n + h_1^{k+1} - 1) \frac{\ell_n}{\bar{\rho}_1^{k+1}} \right. \\ &\quad \left. - x_{n-h_1^k+1}(t - \tau_m) - (i - n + h_1^k - 1) \frac{\ell_n}{\bar{\rho}_1^k} \right| + \int_{t-\tau_m}^t v(R_i(s)) ds \\ &\leq \left| x_{n-h_1^{k+1}+1}(t) - x_{n-h_1^k+1}(t - \tau_m) \right| + (Q + L) \left| \frac{1}{\bar{\rho}_1^{k+1}} - \frac{1}{\bar{\rho}_1^k} \right| \\ &\quad + (h_1^{k+1} - h_1^k) \frac{\ell_n}{\bar{\rho}_1^{k+1}} + \tau_m v_{\max} \\ &\leq \left| x_{n-h_1^{k+1}+1}(t) - x_{n-h_1^{k+1}+1}(t - \tau_m) \right| + \sum_{i=n-h_1^{k+1}+1}^{n-h_1^k} |x_{i+1}(t - \tau_m) - x_i(t - \tau_m)| \\ &\quad + (Q + L) \frac{\tau_m}{\delta^2} \text{Lip}(\bar{\rho}_1) + \frac{\tau_m}{\delta} v_{\max} \rho_{\max} + \tau_m v_{\max} \\ &\leq \left[\frac{Q + L}{\delta^2} \text{Lip}(\bar{\rho}_1) + 2 \frac{v_{\max} \rho_{\max}}{\delta} + 2 v_{\max} \right] \tau_m, \end{aligned} \quad (26)$$

where we have used the minimum principle $R_i(t) \geq \delta$ for all $t \geq 0$ given in Lemma 1, and (twice) the estimate

$$h_1^{k+1} - h_1^k \leq \frac{\tau_m v_{\max}}{\ell_n / \rho_{\max}},$$

which expresses the fact that the total number of particles crossing a given point on a time interval of size τ_m is bounded by the maximum distance covered, i.e., $\tau_m v_{\max}$, divided by the smallest possible distance between two consecutive vehicles, i.e.,

ℓ_n/ρ_{\max} . Finally, by a similar procedure as in (25), we estimate for $i \in \{n - h_1^{k+1} + 1, \dots, n - 1\}$

$$\begin{aligned} & \left| (\bar{\rho}_1^{k+1})^{-1} - R_i(t^-)^{-1} \right| \\ & \leq \left| R_i(t^-)^{-1} - R_i((t - \tau_m)^+)^{-1} \right| + \left| R_i((t - \tau_m)^+)^{-1} - (\bar{\rho}_1^{k+1})^{-1} \right| \\ & \leq \frac{1}{\ell_n} \int_{t-\tau_m}^t \left| v(R_{i+1}(s)) - v(R_i(s)) \right| ds + \left| R_i((t - \tau)^+)^{-1} - (\bar{\rho}_1^{k+1})^{-1} \right| \\ & \leq \frac{v_{\max}}{\ell_n} \tau_m + \left| R_i((t - \tau)^+)^{-1} - (\bar{\rho}_1^{k+1})^{-1} \right|. \end{aligned}$$

Now, the last term on the right-hand side of the above last estimate can be controlled in the case $i \geq n - h_1^k + 1$ by

$$\left| R_i((t - \tau)^+)^{-1} - (\bar{\rho}_1^{k+1})^{-1} \right| = \left| (\bar{\rho}_1^{k+1})^{-1} - (\bar{\rho}_1^k)^{-1} \right| \leq \frac{1}{\delta^2} \text{Lip}(\bar{\rho}_1) \tau_m, \quad (27)$$

while in the case $i < n - h_1^k + 1$ by

$$\begin{aligned} & \left| R_i((t - \tau)^+)^{-1} - (\bar{\rho}_1^{k+1})^{-1} \right| \leq \sum_{j=n-h_1^{k+1}+1}^i \left| R_{j+1}((t - \tau)^+)^{-1} - R_j((t - \tau)^+)^{-1} \right| \\ & \leq (h_1^{k+1} - h_1^k) \frac{\rho_{\max}}{\delta^2} \leq \frac{v_{\max} \rho_{\max}^2}{\ell_n \delta^2} \tau_m. \end{aligned} \quad (28)$$

Hence, substituting (24), (25), (26), (27), and (28) into (23), using $q_n \leq \ell_n$ and the arbitrariness of $t = (k+1)\tau_m$, we can easily find a positive constant $C = C(\delta, \rho_{\max}, v_{\max}, T, \bar{\rho}, \bar{\rho}_0, \bar{\rho}_1) \geq 0$ such that

$$W_1(\rho(t^+, \cdot), \rho(t^-, \cdot)) \leq C \tau_m, \quad t \in (\mathbb{N} + 1) \tau_m. \quad (29)$$

Now, we use the two estimates (22) and (29) to obtain (20). Let $h > 0$ be fixed. Let $t \in [0, T - h]$ and assume for simplicity that $t, t + h \notin \{k\tau_n\}_{k=0}^{m-1}$. We first assume $\tau_m < h$. More precisely, let $t \in (k\tau_m, (k+1)\tau_m)$ and $t + h \in (r\tau_m, (r+1)\tau_m)$ for some $k < r < m$. We have

$$\begin{aligned} W_1(\rho(t+h), \rho(t)) & \leq W_1(\rho(t+h), \rho((r\tau_m)^+)) + W_1(\rho((r\tau_m)^+), \rho((r\tau_m)^-)) \\ & + \sum_{j=k+1}^{r-1} \left[W_1(\rho((j+1)\tau_m)^-, \rho(j\tau_m)^+) + W_1(\rho(j\tau_m)^+, \rho((j\tau_m)^-) \right] \\ & + W_1(\rho(((k+1)\tau_m)^-), \rho(t)) \leq 2C\tau_m + 2C(r-k-1)\tau_m + C\tau_m \\ & \leq C[2(r-k)+1]\tau_m \leq 5Ch, \end{aligned}$$

because by assumption $\tau_m < h$ and $(r - k)\tau_m \leq h + \tau_m \leq 2h$. Since C is some positive constant independent of n, m , and h , we have (H2'), hence (20). Let us now assume $\tau_m \geq h$. In this case, we have by (21) and (29)

$$\begin{aligned}
& \int_0^{T-h} W_1(\rho(t+h), \rho(t)) dt = \int_{(m-1)\tau_m}^{T-h} W_1(\rho(t+h), \rho(t)) dt \\
& + \sum_{k=1}^{m-1} \left[\int_{(k-1)\tau_m}^{k\tau_m-h} W_1(\rho(t+h), \rho(t)) dt + \int_{k\tau_m-h}^{k\tau_m} W_1(\rho(t+h), \rho(t)) dt \right] \\
& \leq \int_{(m-1)\tau_m}^{T-h} W_1(\rho(t+h), \rho(t)) dt + \sum_{k=1}^{m-1} \int_{(k-1)\tau_m}^{k\tau_m-h} W_1(\rho(t+h), \rho(t)) dt \\
& + \sum_{k=1}^{m-1} \left[\int_{k\tau_m-h}^{k\tau_m} \left(W_1(\rho(t+h), \rho((k\tau_m)^+)) + W_1(\rho((k\tau_m)^+), \rho((k\tau_m)^-)) \right. \right. \\
& \quad \left. \left. + W_1(\rho((k\tau_m)^-), \rho(t)) \right) dt \right] \\
& \leq \frac{3}{2}(Q+L)v_{\max}h\tau_m + 3 \sum_{k=1}^{m-1} \left[\frac{3}{2}(Q+L)v_{\max}h\tau_m \right] + \sum_{k=1}^{m-1} [C h \tau_m] \\
& \leq \left[\frac{9}{2}(Q+L)v_{\max} + C \right] T h,
\end{aligned}$$

for some positive constant C independent of n, m , and h . Hence, (20) is proven.

In order to conclude the proof of Theorem 4, we can proceed exactly as in Theorem 3 by using Theorem 1. We observe that condition (H2) of Theorem 1 is used in this case in order to get a uniform continuity estimate in time.

3.3 Convergence to Entropy Solutions

In this subsection, we briefly point out that the scheme introduced in Subsection 3.1 is consistent in some simple cases.

Theorem 5 Assume (I3) and (V1). If $\bar{\rho}_0$ and $\bar{\rho}_1$ are constant and

- either also $\bar{\rho}$ is constant,
- or $f'(\bar{\rho}_0(t)) < 0$ and $f'(\bar{\rho}_1(t)) > 0$ for all $t \geq 0$,

then $(\rho^{n,m})_{n,m}$ converges (up to a subsequence) to the unique entropy solution to the IVP (16) in the sense of Definition 2.

Proof The proof easily follows from the fact that in both cases, the unique entropy solution to (16) on $[0, \tau]$ is the restriction of the solution to the Cauchy problem with

initial condition $\bar{\rho}_0 \mathbf{1}_{(-\infty, 0)} + \bar{\rho} \mathbf{1}_{\Omega} + \bar{\rho}_1 \mathbf{1}_{(1, +\infty)}$. This can be easily seen via a WFT argument (see, e.g., [25] and Remark 5). Hence, one can restart the Cauchy problem on $[\tau, 2\tau]$ with the same construction and proceed iteratively for all times. The above claim proves that for any fixed m , the limit ρ^m of $\rho^{n,m}$ as $n \rightarrow +\infty$ is an entropy solution to (16). The assertion then easily follows by the continuity with respect to the boundary conditions proven in [25, Theorem 2.3.5b].

4 The Hughes Model

In this section, we apply the Hughes model [49] to simulate the evacuation of a one-dimensional corridor $\Omega \doteq (-1, 1)$ ending with two exits. The resulting model is expressed by the following IBVP with Dirichlet boundary conditions

$$\begin{cases} \rho_t - \left[\rho v(\rho) \frac{\phi_x}{|\phi_x|} \right]_x = 0, & x \in \Omega, t > 0, \\ |\phi_x| = c(\rho), & x \in \Omega, t > 0, \\ (\rho, \phi)(t, -1) = (\rho, \phi)(t, 1) = (0, 0), & t > 0, \\ \rho(0, x) = \bar{\rho}(x), & x \in \Omega. \end{cases} \quad (30)$$

We assume that the initial density $\bar{\rho}$ and the velocity map v satisfy (I1) and (V1), respectively, where ρ_{\max} is the maximal crowd density and v_{\max} is the maximal speed of a pedestrian. Let $L \doteq \|\bar{\rho}\|_{L^1(\Omega)}$ and $R \doteq \|\bar{\rho}\|_{L^\infty(\Omega)}$. The maximum principle in [37] shows that ρ never exceeds the range $[0, R]$. We assume also what follows.

There exists a $\hat{\rho} \in (0, \rho_{\max})$ such that for any $\rho \in (0, \rho_{\max}) \setminus \{\hat{\rho}\}$
 $[v(\rho) + \rho v'(\rho)](\hat{\rho} - \rho) > 0$. (V3)

$c: [0, \rho_{\max}] \rightarrow [1, +\infty]$ is \mathbf{C}^2 , $c' \geq 0$, $c'' > 0$, $c(0) = 1$, and $c(R) < +\infty$. (C)

Example 2 In the literature, see [2, 3, 19, 32, 49, 50, 66], the usual choice for the cost function is $c(\rho) \doteq 1/v(\rho)$. In this case, in order to bypass the technical issue of c blowing up at $\rho = \rho_{\max}$, it is assumed that $R \doteq \|\bar{\rho}\|_{L^\infty(\Omega)} \in (0, \rho_{\max})$. This assumption, together with the maximum principle obtained in [37], ensures that the cost computed along any solution of (30) is well defined.

As observed in [2, 3, 37], the differential equations in (30) can be reformulated as

$$\begin{cases} \rho_t + F(t, x, \rho)_x = 0, & x \in \Omega, t > 0, \\ \int_{-1}^{\xi(t)} c(\rho(t, y)) dy = \int_{\xi(t)}^1 c(\rho(t, y)) dy, & x \in \Omega, t > 0, \end{cases} \quad (31)$$

with $F(t, x, \rho) \doteq \text{sign}(x - \xi(t)) f(\rho)$ (see [31] for the details). The form (31) clearly suggests that Hughes' model can be seen as a two-sided LWR model, with the turning point $\xi(t)$ splitting the whole interval Ω into two subintervals. For this reason, under appropriate assumptions that guarantee the presence of a persistent

vacuum region around $\xi(t)$, we can apply the results obtained in Section 2 to (30). The notion of solution in the case of a vacuum region around $t \mapsto \xi(t)$ is as follows

Definition 3 Assume (I1), (V1), (V3), and (C). A map $\rho \in \mathbf{L}^\infty(\mathbb{R}_+ \times \mathbb{R}; [0, R])$ is a (well-separated) *entropy solution* to (30) if

- There exists $\varepsilon > 0$ such that ρ is equal to zero on the open cone

$$\mathcal{C} \doteq \{(t, x) \in \mathbb{R}_+ \times \mathbb{R} : |x - \bar{\xi}| < \varepsilon t\}.$$

- $\rho \mathbf{1}_{(-\infty, \bar{\xi})}$ is the entropy solution to (9) with initial datum $\bar{\rho} \mathbf{1}_{(-\infty, \bar{\xi})}$ in the sense of Definition 1.
- $\rho \mathbf{1}_{(\bar{\xi}, +\infty)}$ is the entropy solution to (9) with initial datum $\bar{\rho} \mathbf{1}_{(\bar{\xi}, +\infty)}$ in the sense of Definition 1.
- The turning curve $\mathcal{T} \doteq \{(t, x) \in \mathbb{R}_+ \times \mathcal{Q} : x = \xi(t)\}$ is continuous and contained in \mathcal{C} . Moreover $(0, \bar{\xi}) \in \mathcal{T}$ and

$$\int_{-1}^{\xi(t)} c(\rho(t, y)) dy = \int_{\bar{\xi}(t)}^1 c(\rho(t, y)) dy, \quad \text{for a.e. } t \geq 0.$$

The next theorem collects the main existence result obtained in [3].

Theorem 6 ([3, Theorem 3]) *If $v(\rho) \doteq 1 - \rho$, $c(\rho) \doteq 1/v(\rho)$ and the initial datum $\bar{\rho} \in \mathbf{BV}(\mathcal{Q}; [0, 1])$ satisfies the estimate $3R + \text{TV}(c(\bar{\rho})) + [c(\bar{\rho}(-1^+)) - c(1/2)]_+ + [c(\bar{\rho}(1^-)) - c(1/2)]_+ < 2$, then there exists an entropy solution to (30) defined globally in time.*

In Section 6.4, we show the numerical simulations of our particle methods in simple Riemann-type initial conditions. We stress here that although the analytical results concerning our deterministic particle method are restricted to cases in which each particle keeps the same direction for all times, the numerical simulations also cover cases with direction switching.

4.1 The Follow-the-Leader Scheme and Main Result

We now introduce our FTL scheme for (30). Assume (I1), (V1), (V3), and (C). Fix $n \in \mathbb{N}$ sufficiently large and set $\ell_n \doteq L/n$. Let $\bar{x}_0^n, \dots, \bar{x}_n^n$ be defined recursively by

$$\begin{cases} \bar{x}_0^n \doteq \min \{\text{supp}(\bar{\rho})\}, \\ \bar{x}_i^n \doteq \inf \left\{ x > \bar{x}_{i-1}^n : \int_{\bar{x}_{i-1}^n}^x \bar{\rho}(y) dy \geq m \right\}, \quad i \in \{1, \dots, n\}. \end{cases}$$

It follows that $-1 \leq \bar{x}_0^n < \bar{x}_1^n < \dots < \bar{x}_{n-1}^n < \bar{x}_n^n \leq 1$ and

$$\int_{\bar{x}_i^n}^{\bar{x}_{i+1}^n} \bar{\rho}(y) dy = \ell_n \leq (\bar{x}_{i+1}^n - \bar{x}_i^n) R, \quad i \in \{0, \dots, n-1\}.$$

We denote the local discrete initial densities

$$\bar{R}_i^n \doteq \frac{\ell_n}{\bar{x}_{i+1}^n - \bar{x}_i^n} \in (0, R], \quad i \in \{0, \dots, n-1\},$$

and introduce the discretized initial density $\bar{\rho}^n : \mathbb{R} \rightarrow [0, \rho_{\max}]$ by

$$\bar{\rho}^n(x) \doteq \sum_{i=0}^{n-1} \bar{R}_i^n \mathbf{1}_{[\bar{x}_i^n, \bar{x}_{i+1}^n)}(x).$$

We implicitly define the initial approximate turning point $\bar{\xi}^n \in \mathcal{Q}$ via the formula

$$\int_{-1}^{\bar{\xi}^n} c(\bar{\rho}^n(y)) dy = \int_{\bar{\xi}^n}^1 c(\bar{\rho}^n(y)) dy.$$

The next step is the definition of the evolving particle scheme. Roughly speaking, $\bar{\xi}^n$ splits the set of particles into left and right particles, the former moving according to a *backward* FTL scheme, the latter according to a *forward* one. By a slight modification of the initial condition, we may always assume that there exists $I_0 \in \{0, \dots, n\}$ such that $\bar{\xi}^n \in (\bar{x}_{I_0}^n, \bar{x}_{I_0+1}^n)$. We then set

$$\begin{cases} \dot{x}_0^n(t) = -v_{\max}, \\ \dot{x}_i^n(t) = -v \left(\frac{\ell_n}{x_i^n(t) - x_{i-1}^n(t)} \right), & i \in \{1, \dots, I_0\}, \\ \dot{x}_i^n(t) = v \left(\frac{\ell_n}{x_{i+1}^n(t) - x_i^n(t)} \right), & i \in \{I_0 + 1, \dots, n-1\}, \\ \dot{x}_n^n(t) = v_{\max}, \\ x_i^n(0) = \bar{x}_i^n, & i \in \{0, \dots, n\}. \end{cases} \quad (32)$$

We consider the corresponding discrete densities

$$\begin{cases} R_i^n(t) \doteq \frac{\ell_n}{x_{i+1}^n(t) - x_i^n(t)}, & i \in \{0, \dots, n-1\} \setminus \{I_0\}, \\ R_i^n(t) \doteq 0, & i \in \{-1, I_0, n\}. \end{cases}$$

Notice that in view of Remark 5, we do not impose any boundary condition in (32), and we follow the movement of each particle whether or not they are in \mathcal{Q} . Moreover, the density has been set to equal zero outside $[x_0^n(t), x_n^n(t)]$ and around the turning point, namely in $[x_{I_0}^n(t), x_{I_0+1}^n(t)]$. The latter in particular is simply due to a consistency with the numerical simulations, in which the computation of the turning point is made simpler in this way. This clearly introduces an error ℓ_n in the

total mass. Finally, the (unique) solution to the system (32) is well defined, and the density $R_{I_0}^n(t)$ is equal to zero until the turning point does not collide with a particle.

The approximated turning point $\xi^n(t)$ is implicitly uniquely defined by

$$\int_{-1}^{\xi^n(t)} c(\rho^n(t, y)) \, dy = \int_{\xi^n(t)}^1 c(\rho^n(t, y)) \, dy,$$

where $\rho^n: \mathbb{R}_+ \times \mathbb{R} \rightarrow [0, \rho_{\max}]$ is the discretized density defined by

$$\rho^n(t, x) \doteq \sum_{i=0}^{n-1} R_i^n(t) \mathbf{1}_{[x_i^n(t), x_{i+1}^n(t))}(x). \quad (33)$$

Clearly $\xi^n(t) \in \Omega$ for all $t \geq 0$ and $\xi^n(0)$ does not necessarily coincide with $\bar{\xi}^n$.

In the next theorem, we state our main result, which deals with a class of *small* initial data in \mathbf{BV} . For further use, we define the function $\Upsilon(\rho) \doteq c(\rho) - c'(\rho)\rho$, which is strictly decreasing in view of assumption (C) above. We then set

$$\begin{aligned} \mathcal{L} &\doteq \text{Lip}[\Upsilon|_{[0, R]}] = \max \{c''(\rho)\rho : \rho \in [0, R]\}, \\ C &\doteq c'(R)R = \max \{c'(\rho)\rho : \rho \in [0, R]\}. \end{aligned}$$

Theorem 7 *Assume (I1), (I2), (V1), (V3) and (C). If the initial datum $\bar{\rho}$ satisfies*

$$R \doteq \|\bar{\rho}\|_{\mathbf{L}^\infty(\Omega)} < \rho_{\max}, \quad \frac{v_{\max}}{2} \left[\mathcal{L} \text{TV}(\bar{\rho}) + 3C \right] < v(R). \quad (34)$$

then there exists a unique entropy solution ρ to (30) in the sense of Definition 3 defined globally in time. Such a solution is obtained as a strong \mathbf{L}^1 -limit of the discrete density ρ^n constructed via the FTL particle system (32).

We omit the proof of Theorem 7, and we defer to [31, Section 2.3] for the details. Let us only remark that the assumption $R < \rho_{\max}$ above is essential in order to have the right-hand side in the inequality (34) strictly positive.

5 The ARZ Model

Consider the Cauchy problem for the ARZ model [9, 69]

$$\begin{cases} \rho_t + (\rho v)_x = 0, & t > 0, x \in \mathbb{R}, \\ (\rho w)_t + (\rho v w)_x = 0, & t > 0, x \in \mathbb{R}, \\ (v, w)(0, x) = (\bar{v}, \bar{w})(x), & x \in \mathbb{R}, \end{cases} \quad (35)$$

where v is the velocity, w is the Lagrangian marker, and (\bar{v}, \bar{w}) is the corresponding initial datum. Moreover, (v, w) belongs to $\mathcal{W} \doteq \{(v, w) \in \bar{\mathbb{R}}_+^2 : v \leq w\}$ and $\rho \doteq p^{-1}(w - v) \geq 0$ is the corresponding density, where $p \in \mathbf{C}^0(\bar{\mathbb{R}}_+; \bar{\mathbb{R}}_+) \cap \mathbf{C}^2(\mathbb{R}_+; \bar{\mathbb{R}}_+)$ satisfies

$$p(0^+) = 0, \quad p'(\rho) > 0 \quad \text{and} \quad 2p'(\rho) + \rho p''(\rho) > 0 \quad \text{for every } \rho > 0. \quad (\text{P})$$

The typical choice is $p(\rho) \doteq \rho^\gamma$, $\gamma > 0$. By definition, we have that the vacuum state $\rho = 0$ corresponds to the half line $\mathcal{W}_0 \doteq \{(v, w)^T \in \mathcal{W} : v = w\}$ and the non-vacuum states $\rho > 0$ to $\mathcal{W}_0^c \doteq \mathcal{W} \setminus \mathcal{W}_0$.

Definition 4 ([5, Definition 2.3] and [6, Definition 2.2]) Let $(\bar{v}, \bar{w}) \in \mathbf{L}^\infty(\mathbb{R}; \mathcal{W})$. We say that a function $(v, w) \in \mathbf{L}^\infty(\bar{\mathbb{R}}_+ \times \mathbb{R}; \mathcal{W}) \cap \mathbf{C}^0(\bar{\mathbb{R}}_+; \mathbf{L}_{\text{loc}}^1(\mathbb{R}; \mathcal{W}))$ is a weak solution of (35) if it satisfies the initial condition $(v(0, x), w(0, x)) = (\bar{v}(x), \bar{w}(x))$ for a.e. $x \in \mathbb{R}$ and for any test function $\phi \in \mathbf{C}_c^\infty(\mathbb{R} \times \mathbb{R})$

$$\iint_{\bar{\mathbb{R}}_+ \times \mathbb{R}} p^{-1}(v, w) (\phi_t + v \phi_x) \begin{pmatrix} 1 \\ w \end{pmatrix} dx dt = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We refer to [40] for the existence of solutions to (35) away from vacuum and to [43] for the existence with vacuum. Let us briefly recall the main properties of the solutions to (35). If the initial density $\bar{\rho} \doteq p^{-1}(\bar{w} - \bar{v})$ has compact support, then the support of ρ has finite speed of propagation. The maximum principle holds true in the Riemann invariant coordinates (v, w) , but not in the conserved variables $(\rho, \rho w)$ as a consequence of hysteresis processes. Moreover, the total space occupied by the vehicles is time independent: $\int_{\mathbb{R}} \rho(t, x) dx = \|\bar{\rho}\|_{\mathbf{L}^1(\mathbb{R})}$ for all $t \geq 0$.

5.1 The Follow-the-Leader Scheme and Main Result

We introduce our atomization scheme for the Cauchy problem (35). Let $(\bar{v}, \bar{w}) \in \mathbf{BV}(\mathbb{R}; \mathcal{W})$ be such that $\bar{\rho} \doteq p^{-1}(\bar{w} - \bar{v})$ belongs to $\mathbf{L}^1(\mathbb{R})$ and $\bar{\rho}$ is compactly supported. Denote by $\bar{x}_{\min} < \bar{x}_{\max}$ the extremal points of the convex hull of the compact support of $\bar{\rho}$, namely

$$\bigcap_{[a,b] \supseteq \text{supp}(\bar{\rho})} [a, b] = [\bar{x}_{\min}, \bar{x}_{\max}].$$

Fix $n \in \mathbb{N}$ sufficiently large. Let $L \doteq \|\bar{\rho}\|_{\mathbf{L}^1(\mathbb{R})} > 0$ and $\ell_n \doteq L/n$. Set recursively

$$\begin{cases} \bar{x}_0^n \doteq \bar{x}_{\min}, \\ \bar{x}_i^n \doteq \sup \left\{ x \in \mathbb{R} : \int_{\bar{x}_{i-1}^n}^x \bar{\rho}(x) dx < \ell_n \right\}, \quad i \in \{1, \dots, n\}. \end{cases} \quad (36)$$

It is easily seen that $\bar{x}_i^n = \bar{x}_{\max}$ for all $i = 0, \dots, n$. We approximate then \bar{w} by taking

$$\bar{w}_i^n \doteq \operatorname{ess\,sup}_{[\bar{x}_i^n, \bar{x}_{i+1}^n]}(\bar{w}), \quad i \in \{0, \dots, n-1\}. \quad (37)$$

We have then

$$\ell_n = \int_{\bar{x}_i^n}^{\bar{x}_{i+1}^n} \bar{\rho}(x) dx \leq (\bar{x}_{i+1}^n - \bar{x}_i^n) \rho_{i,\max}^n, \quad i \in \{0, \dots, n-1\},$$

with $\rho_{i,\max}^n \doteq p^{-1}(\bar{w}_i^n)$. We take the values $\bar{x}_0^n, \dots, \bar{x}_n^n$ as the initial positions of the $(n+1)$ particles in the n -depending FTL model

$$\begin{cases} x_n^n(t) = \bar{x}_{\max} + \bar{w}_{n-1}^n t, \\ \dot{x}_i^n(t) = v_i^n \left(\frac{\ell_n}{x_{i+1}^n(t) - x_i^n(t)} \right), \quad i \in \{0, \dots, n-1\}, \\ x_i^n(0) = \bar{x}_i^n, \quad i \in \{0, \dots, n\}, \end{cases} \quad (38)$$

where

$$v_i^n(\rho) \doteq \bar{w}_i^n - p(\rho), \quad i \in \{0, \dots, n-1\}. \quad (39)$$

The quantity $\bar{w}_i^n = v_i^n(0)$ is the maximum possible velocity allowed for the i th vehicle. Clearly, only the leading vehicle x_n^n reaches its maximal velocity, as the vacuum state is achieved only ahead of x_n^n . The existence of a global solution to (38) follows from [29, Lemma 2.3], which generalizes the discrete maximum principle of Lemma 1. Finally, since v_i^n is decreasing, and its argument $\ell_n/[x_{i+1}^n(t) - x_i^n(t)]$ is always bounded above by $\rho_{i,\max}^n$, we have $x_0^n(t) \geq \bar{x}_{\min} + v_0(R_0^n) t = \bar{x}_{\min}$. By introducing in (38)

$$R_i^n(t) \doteq \frac{\ell_n}{x_{i+1}^n(t) - x_i^n(t)}, \quad i \in \{0, \dots, n-1\}, \quad (40)$$

we obtain

$$\begin{cases} \dot{R}_{n-1}^n = -\frac{(R_{n-1}^n)^2}{\ell_n} p(R_{n-1}^n), \\ \dot{R}_i^n = -\frac{(R_i^n)^2}{\ell_n} [v_{i+1}^n(R_{i+1}^n) - v_i^n(R_i^n)], \quad i \in \{0, \dots, n-2\}, \\ R_i^n(0) = \bar{R}_i^n \doteq \frac{\ell_n}{\bar{x}_{i+1}^n - \bar{x}_i^n}, \quad i \in \{0, \dots, n-1\}. \end{cases} \quad (41)$$

Observe that $\ell_n/[\bar{x}_{\max} - \bar{x}_{\min} + \bar{w}_{n-1}^n t] \leq R_i^n(t) \leq \rho_{i,\max}^n$ for all $t \geq 0$ in view of the discrete maximum principle. The quantity R_i^n can be seen as a discrete version of the density ρ in Lagrangian coordinates, and (41) is the discrete Lagrangian version of the Cauchy problem (35).

Define the piecewise constant (with respect to x) Lagrangian marker

$$W^n(t, x) \doteq \begin{cases} \bar{w}_0^n & \text{if } x \in (-\infty, x_0^n(t)), \\ \bar{w}_i^n & \text{if } x \in [x_i^n(t), x_{i+1}^n(t)], \quad i \in \{0, \dots, n-1\}, \\ \bar{w}_{n-1}^n & \text{if } x \in [x_n^n(t), +\infty), \end{cases} \quad (42)$$

and the piecewise constant (with respect to x) velocity

$$V^n(t, x) \doteq \begin{cases} \bar{v}_0^n & \text{if } x \in (-\infty, x_0^n(t)), \\ v_i^n(R_i^n(t)) & \text{if } x \in [x_i^n(t), x_{i+1}^n(t)], \quad i \in \{0, \dots, n-1\}, \\ \bar{w}_{n-1}^n & \text{if } x \in [x_n^n(t), +\infty). \end{cases} \quad (43)$$

We are now ready to state the main result proved in [29].

Theorem 8 Assume (P). Let $(\bar{v}, \bar{w}) \in \mathbf{BV}(\mathbb{R}; \mathcal{W})$ be such that $\bar{\rho} \doteq p^{-1}(\bar{w} - \bar{v})$ is compactly supported and belongs to $\mathbf{L}^1(\mathbb{R})$. Fix $n \in \mathbb{N}$ sufficiently large and let $\ell_n \doteq L/n$, with $L \doteq \|\bar{\rho}\|_{\mathbf{L}^1(\mathbb{R})}$. Let $\bar{x}_0^n < \dots < \bar{x}_n^n$ be the atomization constructed in (36). Let $x_0^n(t), \dots, x_n^n(t)$ be the solution to the FTL system (38). Let $\bar{w}_0^n, \dots, \bar{w}_{n-1}^n$ be given by (37). Set W^n and V^n as in (42) and (43) respectively, where v_i^n and R_i^n are defined by (39) and (40) respectively. Then, $(V^n, W^n)_n$ converges (up to a subsequence) in $\mathbf{L}^1_{\text{loc}}(\bar{\mathbb{R}}_+ \times \mathbb{R}; \mathcal{W})$ as $n \rightarrow +\infty$ to a weak solution of the Cauchy problem (35) with initial datum (\bar{v}, \bar{w}) in the sense of Definition 4.

We omit the proof of Theorem 8, and we defer to [29, Theorem 3.2] for the details. For completeness, we point out that the corresponding discrete density is

$$\rho^n(t, x) \doteq p^{-1}(W^n(t, x) - V^n(t, x)) = \sum_{i=1}^{n-1} R_i^n(t) \chi_{[x_i^n(t), x_{i+1}^n(t)]}(x).$$

6 Numerical Simulations

In this section, we present numerical simulations for the particle method described above. We compare the numerical simulations with the exact solutions obtained by the method of characteristics and that with the Godunov method. The particle system is solved by using the Runge–Kutta MATLAB solver ODE23, with the initial mesh size determined by the total number of particles N and the initial density values.

6.1 The Cauchy Problem for the LWR Equation

We first furnish one example for the Cauchy problem for the LWR equation (9) with flux given by $f(\rho) \doteq \rho(1 - \rho)$. In Fig. 1, we take $N = 200$, final time $T = 0.5$, and initial datum

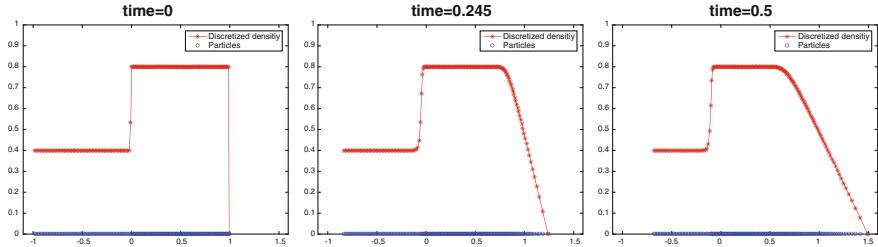


Fig. 1 The evolution of ρ^n given by (13) and corresponding to the initial datum (44). The circles in the bottom (in blue in the electronic version) denote particle location, while the stars in the top (in red in the electronic version) denote the computed density.

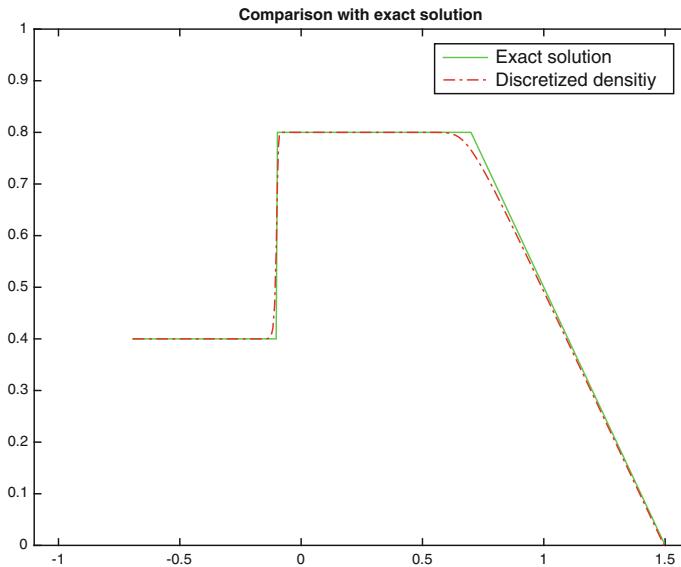


Fig. 2 Comparison between the exact solution (continuous green line in the electronic version) and ρ^n ('— · —' in red in the electronic version) for $N = 400$ and initial datum (44).

$$\bar{\rho}(x) = \begin{cases} 0.4 & \text{if } -1 \leq x \leq 0, \\ 0.8 & \text{if } 0 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (44)$$

In Fig. 2, we compare the result of the simulation with $N = 400$ particles and final time $t = 0.5$ with exact solutions.

6.2 The Cauchy–Dirichlet Problem for the LWR Equation

More interesting situations can be illustrated in the case of LWR with Dirichlet boundary conditions (16) (see Figures 3 and 4). As pointed out in Section 3, the atomization algorithm introduces artificial queuing particles for miming the left boundary condition. In $x < 0$, we arrange N queuing particles (the ones that are going to enter in the domain at time T), with N given by (17). Again, we take $f(\rho) \doteq \rho(1 - \rho)$. For $N = 100$ particles, we consider, according to the notation in Section 3, $\bar{\rho}(x) = 0.2$ in Figures 3 and 4 with left boundary condition $\bar{\rho}_0 = 0.4$ and right boundary conditions $\bar{\rho}_1 = 0$ and $\bar{\rho}_1 = 1$, respectively. In Fig. 5, we set

$$\bar{\rho}(x) = \begin{cases} 0.8 & \text{if } x \in [0, 0.5], \\ 0.1 & \text{if } x \in (0.5, 1], \end{cases} \quad \bar{\rho}_0 = 0.3, \quad \bar{\rho}_1 = 0.1. \quad (45)$$

The latter example is chosen in such way that the actual entropy solution does not match the solution obtained without reupdating the boundary condition at $x = 0$ at every time step. A comparison between discretized densities and numerical solutions obtained via Godunov scheme is also plotted in Figures 3 and 4. We set the spatial discretization according to the number of particles N ; the time step is the same for

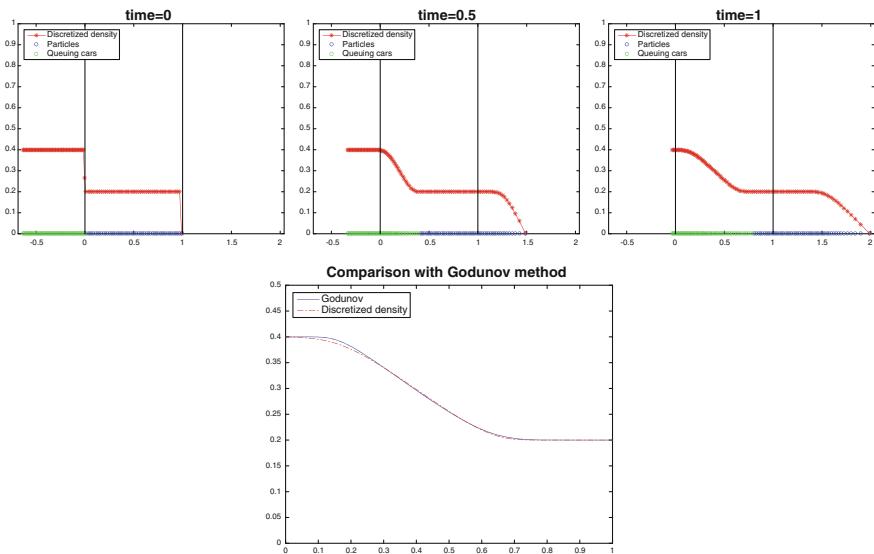


Fig. 3 The circles in the bottom are now divided into two groups: particles that are initially inside the domain (blue in the electronic version) and the queuing particles (green in the electronic version). The star-shaped line in the top (in red in the electronic version) denotes the computed density. Vertical black lines denote the boundary of $[0, 1]$. The initial–boundary setting produces two rarefaction waves both traveling from the left to the right. In the bottom row, we present a comparison with Godunov method.

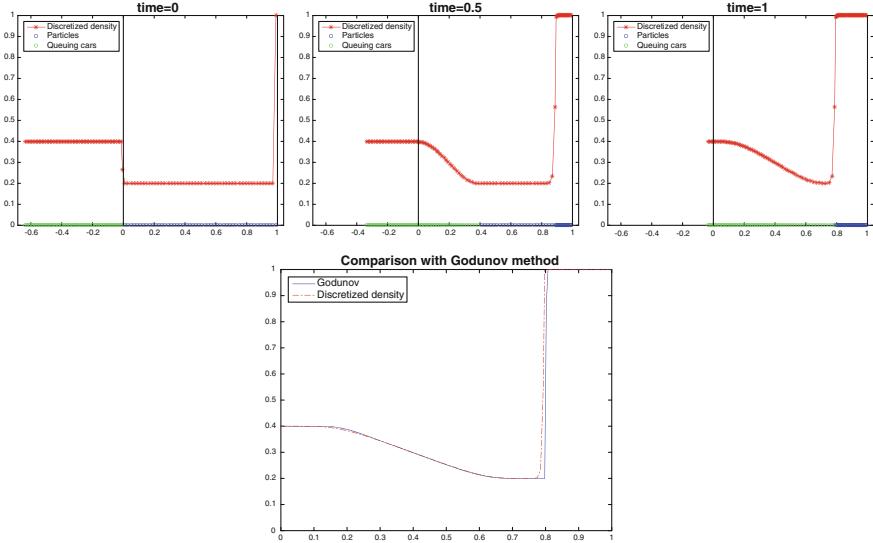


Fig. 4 In this situation, a shock wave travels backward. The notation is similar to Figure 3

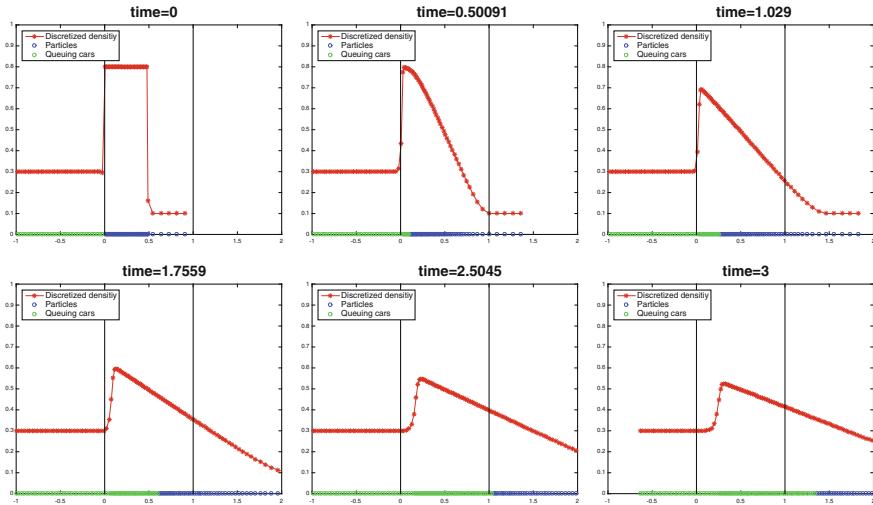


Fig. 5 Simulation for initial–boundary data given in (45).

both methods and is selected so that the CFL condition for the Godunov method holds. Empirically, the observed time step restriction for the FTL method is much less severe than for the Godunov method applied to the Eulerian description of the flow.

Time-dependent piecewise constant boundary data are considered in Fig. 6, where for $N = 400$, we set

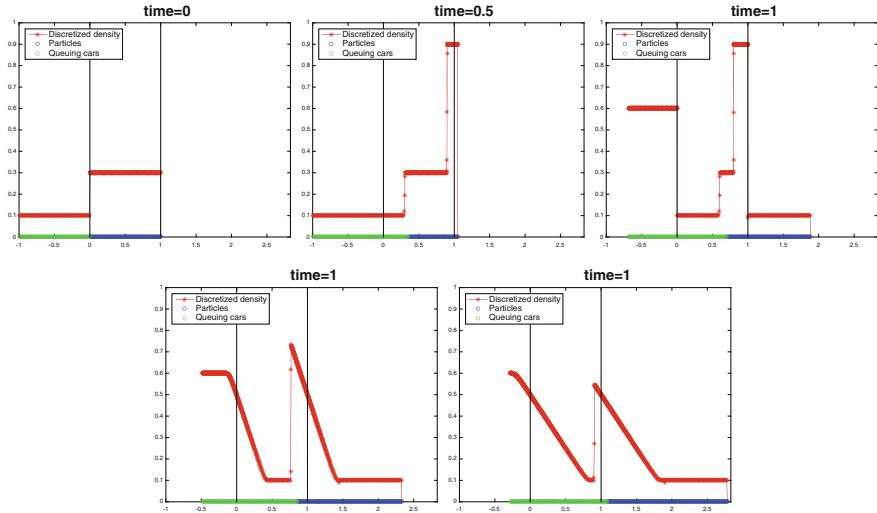


Fig. 6 Simulation for initial-boundary data given in (46).

$$\bar{\rho}(x) = 0.3, \quad \bar{\rho}_0(t) = \begin{cases} 0.1 & \text{if } t \in [0, 1], \\ 0.6 & \text{if } t \in (1, 2], \end{cases} \quad \bar{\rho}_1(t) = \begin{cases} 0.9 & \text{if } t \in [0, 1], \\ 0.1 & \text{if } t \in (1, 2]. \end{cases} \quad (46)$$

Using these conditions, one can built the exact solutions at time $T = 2$

$$\rho_{ex}(2, x) = \begin{cases} 0.5(1 - x) & \text{if } x \in [0, 0.8], \\ 0.1 & \text{if } x \in (0.8, 0.2(9 - 2\sqrt{5})], \\ 0.5(2 - x) & \text{if } x \in (0.2(9 - 2\sqrt{5}), 1]. \end{cases} \quad (47)$$

A comparison with the exact solution ρ_{ex} is given in Fig. 7.

6.3 The ARZ Model

For the ARZ model (35), we consider two examples of Riemann problem. The first one coincides with that one shown in [22, Section 4] and is used to check the ability of the scheme to deal with contact discontinuities. The second one is the example given in [8, Section 5] and is used to check the ability of the scheme to deal with vacuum. The qualitative results corresponding to $N = 200$ and final time $T = 0.2$ for the Test 1 and $T = 1$ for Test 4 are presented in Fig. 8.

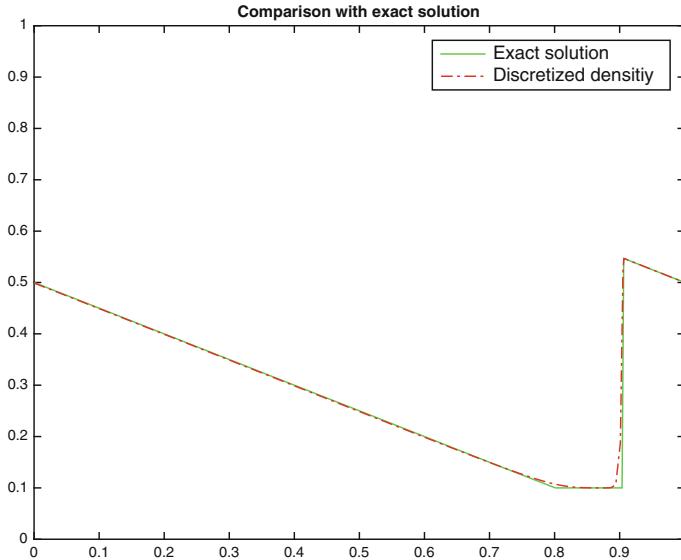


Fig. 7 Comparison between the approximate and the exact solution given in (47) to the IVP (16) with data given in (46).

6.4 The Hughes Model for Pedestrian Movements

In this section, we compare our discrete density for the Hughes model (30) with approximate solutions obtained via Godunov scheme. About the boundary conditions, as pointed out in Section 4.1, we do not impose any boundary condition in the particle method. For the Godunov method, we create two extra ghost cells, one just at the left of -1 and one just at the right of 1 , setting $\rho = 0$ in those cells, to mimic ‘perfect exits.’ In the example reported, the choice for the cost function is $c(\rho) \doteq 1/v(\rho)$, with $v(\rho) \doteq 1 - \rho$, and we show time evolution of the discrete density ρ^n given by (33) in the domain $(-1, 1)$. In order to compare our method with the tests performed in [32, 42], in Fig. 9, we consider the three-step initial condition

$$\bar{\rho}(x) = \begin{cases} 0.8 & \text{if } -0.8 < x \leq -0.5, \\ 0.6 & \text{if } -0.3 < x \leq 0.3, \\ 0.9 & \text{if } 0.4 < x \leq 0.75, \\ 0 & \text{otherwise.} \end{cases} \quad (48)$$

As shown in Figures 9 and 10, this example exhibits the typical *mass transfer* phenomenon occurring when the turning point $\xi(t)$ is not surrounded by a vacuum region. In such a case, particles are crossing $\xi(t)$, and a *non-classical shock* starts from $\xi(t)$, see [2, Remark 5]. In the example, we set $N = 200$ and plot the particle positions and the discrete densities. In Fig. 10, we compare the particle method and a classical

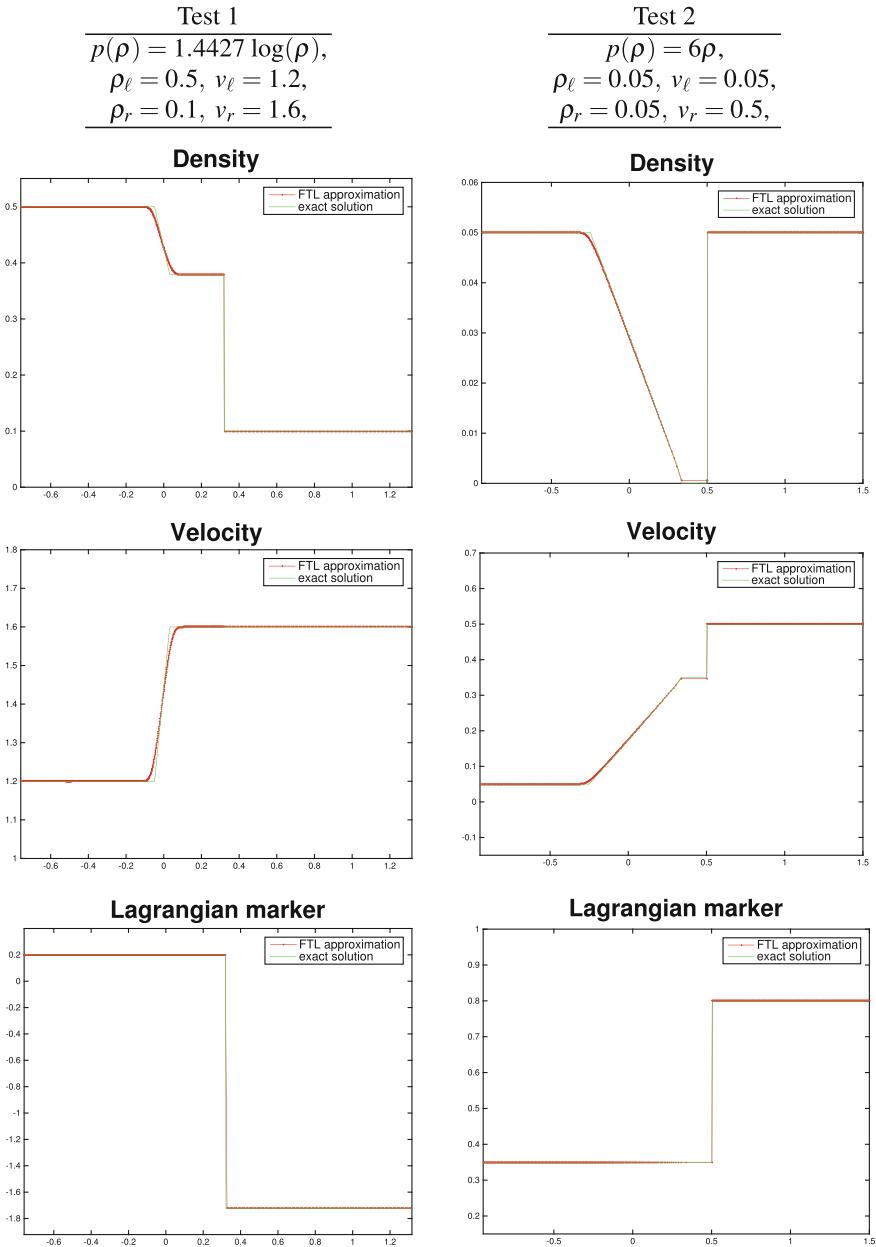


Fig. 8 Left column for Test 1 and right column for Test 2, with $N = 200$.

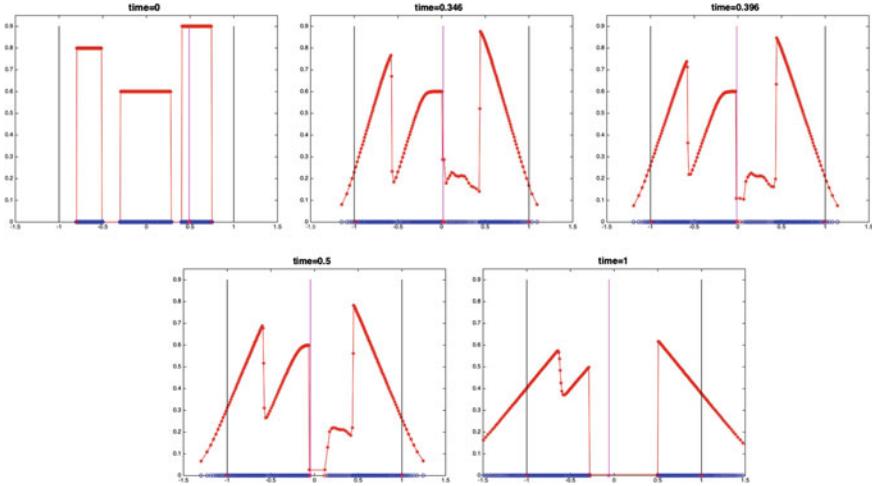


Fig. 9 Evolution of ρ^n (in red) with initial data given in (48). The blue dots represent particles positions. The magenta vertical line is the turning point. In second and third snapshot, we see mass transfer across the turning point and non-classical shock.

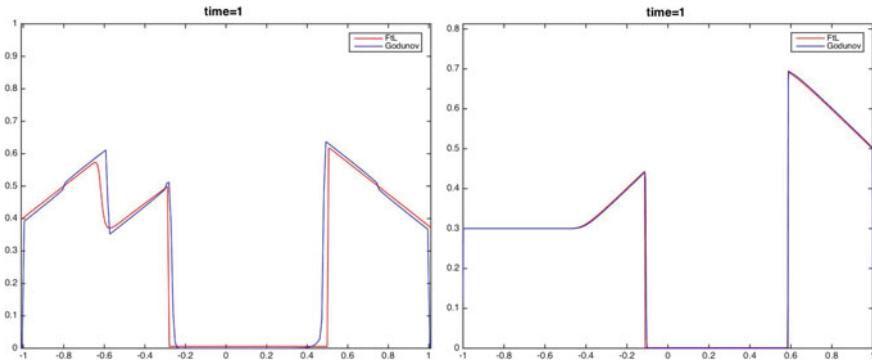


Fig. 10 Comparison between the FTL (in red) and the Godunov (in blue) schemes for the Hughes model (30). On the left for the initial datum given in (48), $N = 100$ and 500 time iterations. On the right for the initial datum $\bar{\rho} \doteq 0.3 \mathbf{1}_{[-1,0]} + 0.7 \mathbf{1}_{(0,1]}$, $N = 1000$ and 1500 time iterations.

Godunov scheme. It is evident that the two methods, though conceptually different, produce approximate solutions are in a good agreement.

Acknowledgements MDF and MDR are supported by the GNAMPA (Italian group of Analysis, Probability, and Applications) project *Geometric and qualitative properties of solutions to elliptic and parabolic equations*. SF and MDR are supported by the GNAMPA (Italian group of Analysis, Probability, and Applications) project *Analisi e stabilità per modelli di equazioni alle derivate parziali nella matematica applicata*. GR was partially supported by ITN-ETN Marie Curie Actions ModCompShock—‘Modeling and Computation of Shocks and Interfaces.’

References

1. D. Amadori and R. M. Colombo. Continuous dependence for 2×2 conservation laws with boundary. *J. Differential Equations*, 138(2):229–266, 1997.
2. D. Amadori and M. Di Francesco. The one-dimensional Hughes model for pedestrian flow: Riemann-type solutions. *Acta Math. Sci. Ser. B Engl. Ed.*, 32(1):259–280, 2012.
3. D. Amadori, P. Goatin, and M. D. Rosini. Existence results for Hughes’ model for pedestrian flows. *J. Math. Anal. Appl.*, 420(1):387–406, 2014.
4. L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. 2nd ed. Lectures in Mathematics, ETH Zürich. Basel: Birkhäuser, 2008.
5. B. Andreianov, C. Donadello, U. Razafison, J. Y. Rolland, and M. D. Rosini. Solutions of the Aw-Rascle-Zhang system with point constraints. *Networks and Heterogeneous Media*, 11(1):29–47, 2016.
6. B. Andreianov, C. Donadello, and M. D. Rosini. A second-order model for vehicular traffics with local point constraints on the flow. *Mathematical Models and Methods in Applied Sciences*, 26(04):751–802, 2016.
7. J.-P. Aubin. Macroscopic traffic models: Shifting from densities to ‘celerities’. *Applied Mathematics and Computation*, 217(3):963–971, 2010.
8. A. Aw, A. Klar, T. Materne, and M. Rascle. Derivation of continuum traffic flow models from microscopic Follow-the-Leader models. *SIAM Journal on Applied Mathematics*, 63(1):259–278, 2002.
9. A. Aw and M. Rascle. Resurrection of “second order” models of traffic flow. *SIAM J. Appl. Math.*, 60(3):916–938 (electronic), 2000.
10. C. Bardos, A. Y. le Roux, and J.-C. Nédélec. First order quasilinear equations with boundary conditions. *Comm. Partial Differential Equations*, 4(9):1017–1034, 1979.
11. N. Bellomo and A. Bellouquid. On the modeling of crowd dynamics: looking at the beautiful shapes of swarms. *Networks and Heterogeneous Media*, 6:383–399, 2011.
12. N. Bellomo, M. Delitala, and V. Coscia. On the mathematical theory of vehicular traffic flow. I. Fluid dynamic and kinetic modelling. *Math. Models Methods Appl. Sci.*, 12(12):1801–1843, 2002.
13. N. Bellomo and C. Dogbe. On the modeling of traffic and crowds: a survey of models, speculations, and perspectives. *SIAM Rev.*, 53(3):409–463, 2011.
14. F. Berthelin, P. Degond, M. Delitala, and M. Rascle. A model for the formation and evolution of traffic jams. *Arch. Ration. Mech. Anal.*, 187(2):185–220, 2008.
15. F. Bolley, Y. Brenier, and G. Loeper. Contractive metrics for scalar conservation laws. *J. Hyperbolic Differ. Equ.*, 2(1):91–107, 2005.
16. Y. Brenier and E. Grenier. Sticky particles and scalar conservation laws. *SIAM J. Numer. Anal.*, 35(6):2317–2328 (electronic), 1998.
17. A. Bressan. Global solutions of systems of conservation laws by wave-front tracking. *J. Math. Anal. Appl.*, 170(2):414–432, 1992.
18. A. Bressan. *Hyperbolic systems of conservation laws*, volume 20 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2000. The one-dimensional Cauchy problem.
19. M. Burger, M. Di Francesco, P. A. Markowich, and M.-T. Wolfram. Mean field games with nonlinear mobilities in pedestrian dynamics. *Discrete Contin. Dyn. Syst. Ser. B*, 19(5):1311–1333, 2014.
20. J. A. Carrillo, M. Di Francesco, and C. Lattanzio. Contractivity of Wasserstein metrics and asymptotic profiles for scalar conservation laws. *J. Differential Equations*, 231(2):425–458, 2006.
21. J. A. Carrillo, S. Martin, and M.-T. Wolfram. An improved version of the Hughes model for pedestrian flow. *Mathematical Models and Methods in Applied Sciences*, 26(04):671–697, 2016.
22. C. Chalons and P. Goatin. Transport-equilibrium schemes for computing contact discontinuities in traffic flow modeling. *Commun. Math. Sci.*, 5(3):533–551, 09 2007.

23. G.-Q. Chen and M. Rascle. Initial layers and uniqueness of weak entropy solutions to hyperbolic conservation laws. *Arch. Ration. Mech. Anal.*, 153(3):205–220, 2000.
24. R. M. Colombo and A. Marson. A Hölder continuous ODE related to traffic flow. *Proc. Roy. Soc. Edinburgh Sect. A*, 133(4):759–772, 2003.
25. R. M. Colombo and M. D. Rosini. Well posedness of balance laws with boundary. *J. Math. Anal. Appl.*, 311(2):683–702, 2005.
26. R. M. Colombo and E. Rossi. On the micro-macro limit in traffic flow. *Rend. Semin. Mat. Univ. Padova*, 131:217–235, 2014.
27. C. M. Dafermos. Polygonal approximations of solutions of the initial value problem for a conservation law. *J. Math. Anal. Appl.*, 38:33–41, 1972.
28. C. F. Daganzo. A variational formulation of kinematic waves: basic theory and complex boundary conditions. *Transportation Research Part B: Methodological*, 39(2):187–196, 2005.
29. M. Di Francesco, S. Fagioli, and M. D. Rosini. Many particle approximation for the Aw-Rascle-Zhang second order model for vehicular traffic. *Mathematical Biosciences and Engineering (MBE)*, 14(1), February 2017 (online).
30. M. Di Francesco, S. Fagioli, and M. D. Rosini. Deterministic particle approximation of scalar conservation laws. *arXiv preprint arXiv:1602.06153*, 2016.
31. M. Di Francesco, S. Fagioli, M. D. Rosini, and G. Russo. Deterministic particle approximation of the Hughes model in one space dimension. *Kinetic and Related Models*, 10(1):215–237, 2017.
32. M. Di Francesco, P. A. Markowich, J.-F. Pietschmann, and M.-T. Wolfram. On the Hughes' model for pedestrian flow: the one-dimensional case. *J. Differential Equations*, 250(3):1334–1362, 2011.
33. M. Di Francesco and M. D. Rosini. Rigorous derivation of nonlinear scalar conservation laws from Follow-the-Leader type models via many particle limit. *Archive for Rational Mechanics and Analysis*, 217(3):831–871, 2015.
34. R. J. DiPerna. Global existence of solutions to nonlinear hyperbolic systems of conservation laws. *J. Differential Equations*, 20(1):187–212, 1976.
35. R. L. Dobrušin. Vlasov equations. *Funktional. Anal. i Prilozhen.*, 13(2):48–58, 96, 1979.
36. F. Dubois and P. LeFloch. Boundary conditions for nonlinear hyperbolic systems of conservation laws. *J. Differential Equations*, 71(1):93–122, 1988.
37. N. El-Khatib, P. Goatin, and M. D. Rosini. On entropy weak solutions of Hughes' model for pedestrian motion. *Z. Angew. Math. Phys.*, 64(2):223–251, 2013.
38. P. A. Ferrari. Shock fluctuations in asymmetric simple exclusion. *Probab. Theory Related Fields*, 91(1):81–101, 1992.
39. P. L. Ferrari and P. Nejjar. Shock fluctuations in flat TASEP under critical scaling. *J. Stat. Phys.*, 160(4):985–1004, 2015.
40. R. E. Ferreira and C. I. Kondo. Glimm method and wave-front tracking for the Aw-Rascle traffic flow model. *Far East J. Math. Sci.*, 43:203–233, 2010.
41. J. Glimm. Solutions in the large for nonlinear hyperbolic systems of equations. *Comm. Pure Appl. Math.*, 18:697–715, 1965.
42. P. Goatin and M. Mimault. The wave-front tracking algorithm for Hughes' model of pedestrian motion. *SIAM J. Sci. Comput.*, 35(3):B606–B622, 2013.
43. M. Godvik and H. Hanche-Olsen. Existence of solutions for the Aw-Rascle traffic flow model with vacuum. *Journal of Hyperbolic Differential Equations*, 05(01):45–63, 2008.
44. L. Gosse and G. Toscani. Identification of asymptotic decay to self-similarity for one-dimensional filtration equations. *SIAM J. Numer. Anal.*, 43(6):2590–2606 (electronic), 2006.
45. H. Greenberg. An analysis of traffic flow. *Operations Research*, 7(1):79–85, 1959.
46. B. Greenshields. A study of traffic capacity. *Proceedings of the Highway Research Board*, 14:448–477, 1935.
47. D. Hoff. The Sharp Form of Oleinik's Entropy Condition in Several Space Variables. *Transactions of the American Mathematical Society*, 276(2):707–714, 1983.
48. H. Holden and N. H. Risebro. *Front tracking for hyperbolic conservation laws*, volume 152. Springer, 2015.

49. R. L. Hughes. A continuum theory for the flow of pedestrians. *Transportation Research Part B: Methodological*, 36(6):507–535, 2002.
50. R. L. Hughes. The flow of human crowds. In *Annual review of fluid mechanics, Vol. 35*, volume 35 of *Annu. Rev. Fluid Mech.*, pages 169–182. Annual Reviews, Palo Alto, CA, 2003.
51. C. Kipnis and C. Landim. *Scaling limits of interacting particle systems*, volume 320 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1999.
52. S. N. Kruzhkov. First order quasilinear equations with several independent variables. *Mat. Sb. (N.S.)*, 81 (123):228–255, 1970.
53. M. J. Lighthill and G. B. Whitham. On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proc. Roy. Soc. London. Ser. A.*, 229:317–345, 1955.
54. P.-L. Lions, B. Perthame, and E. Tadmor. A kinetic formulation of multidimensional scalar conservation laws and related equations. *J. American Math. Society*, 7:169–191, 1994.
55. D. Matthes and H. Osberger. Convergence of a variational Lagrangian scheme for a nonlinear drift diffusion equation. *ESAIM Math. Model. Numer. Anal.*, 48(3):697–726, 2014.
56. C. B. Morrey, Jr. On the derivation of the equations of hydrodynamics from statistical mechanics. *Comm. Pure Appl. Math.*, 8:279–326, 1955.
57. H. Neunzert, A. Klar, and J. Struckmeier. Particle methods: theory and applications. In *ICIAM 95 (Hamburg, 1995)*, volume 87 of *Math. Res.*, pages 281–306. Akademie Verlag, Berlin, 1996.
58. G. F. Newell. A simplified theory of kinematic waves in highway traffic. *Transportation Research Part B: Methodological*, 27(4):281–313, 1993.
59. O. A. Oleinik. Discontinuous solutions of nonlinear differential equations. *Amer. Math. Soc. Transl. (2)*, 26:95–172, 1963.
60. L. Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Phys. Rev. (2)*, 65:117–149, 1944.
61. B. Piccoli and A. Tosin. Vehicular traffic: A review of continuum mathematical models. In R. A. Meyers, editor, *Encyclopedia of Complexity and Systems Science*. Springer New York, 2009.
62. L. A. Pipes. Car following models and the fundamental diagram of road traffic. *Transp. Res.*, 1:21–29, 1967.
63. P. I. Richards. Shock waves on the highway. *OPERATIONS RESEARCH*, 4(1):42–51, 1956.
64. M. D. Rosini. *Macroscopic models for vehicular flows and crowd dynamics: theory and applications*. Understanding Complex Systems. Springer, Heidelberg, 2013.
65. G. Russo. Deterministic diffusion of particles. *Comm. on Pure and Applied Mathematics*, 43:697–733, 1990.
66. M. Twarogowska, P. Goatin, and R. Duvigneau. Macroscopic modeling and simulations of room evacuation. *Appl. Math. Model.*, 38(24):5781–5795, 2014.
67. R. T. Underwood. Speed, volume, and density relationship. In *Quality and theory of traffic flow: a symposium*, pages 141–188. Greenshields, B.D. and Bureau of Highway Traffic, Yale University, 1961.
68. C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
69. H. M. Zhang. A non-equilibrium traffic model devoid of gas-like behavior. *Transportation Research Part B: Methodological*, 36(3):275–290, 2002.

Mean Field Limit for Stochastic Particle Systems

Pierre-Emmanuel Jabin and Zhenfu Wang

Abstract We review some classical and more recent results for the derivation of mean field equations from systems of many particles, focusing on the stochastic case where a large system of SDEs leads to a McKean–Vlasov PDE as the number N of particles goes to infinity. Classical mean field limit results require that the interaction kernel be essentially Lipschitz. To handle more singular interaction kernels is a long-standing and challenging question but which has had some recent successes.

1 Introduction

Large systems of interacting particles are now fairly ubiquitous. The corresponding microscopic models are usually conceptually simple, based for instance on Newton’s second law. However, they are analytically and computationally complicated since the number N of particles is very large (with N in the range of $10^{20}–10^{25}$ for typical physical settings).

Understanding how this complexity can be reduced is a challenging but critical question with potentially deep impact in various fields and a wide range of applications: in physics where particles can represent ions and electrons in plasmas or molecules in a fluid and even galaxies in some cosmological models; in biosciences where they typically model microorganisms (cells or bacteria); and in economics or social sciences where particles are individual “agents.”

P.-E. Jabin is partially supported by NSF Grant 1312142 and by NSF Grant RNMS (Ki-Net) 1107444.

Z. Wang is supported by NSF Grant 1312142.

P.-E. Jabin (✉) · Z. Wang
CSCAMM and Dept. of Mathematics, University of Maryland,
College Park, MD 20742, USA
e-mail: pjabin@cscamm.umd.edu

Z. Wang
e-mail: zwang423@math.umd.edu

The classical strategy to reduce this complexity is to derive a mesoscopic or macroscopic system, *i.e.*, a continuous description of the dynamics where the information is embedded in densities typically solving nonlinear PDEs.

The idea of such a kinetic description of large systems of particles goes back to the original derivation of statistical mechanics and the works of Maxwell and Boltzmann on what is now called the Boltzmann equation which describes the evolution of dilute gases.

We consider here a different (and in several respects easier) setting: The mean field scaling is in a collisionless regime, meaning that collisions seldom occur and particles interact with each other at long range.

The first such mean field equation was introduced in galactic dynamics by Jeans in 1915 [49]. The Vlasov equation was introduced in plasma physics by Vlasov in [75, 76] as the mean field equation for large particle systems of ions or electrons interacting through the Coulomb force, ignoring the effect of collisions.

The rigorous derivation of the Vlasov equation with the Coulomb or Newtonian potential from Newton dynamics is still a major open question in the topic. For some recent progress in that direction, we refer to [42, 54, 55]. However, we focus here on the stochastic case and refer to [35, 46] for a review of the mean field limit for deterministic systems.

In the rest of this introduction, we present some of the classical models that one typically considers.

1.1 Classical Second-Order Dynamics

The most classical model is the Newton dynamics for N indistinguishable point particles driven by two-body interaction forces and Brownian motions. Denote by $X_i \in \Gamma$ and $V_i \in \mathbb{R}^d$ the position and velocity of particle number i . The evolution of the system is given by the following SDEs:

$$dX_i = V_i dt, \quad dV_i = \frac{1}{N} \sum_{j \neq i} K(X_i - X_j) dt + \sqrt{2\sigma} dW_t^i, \quad (1)$$

where $i = 1, 2, \dots, N$. The W^i are N independent Brownian motions or Wiener processes, which may model various types of random phenomena, for instance random collisions against a given background. If $\sigma = 0$, the system (1) reduces to the classical deterministic Newton dynamics. We always assume here that $\sigma > 0$, but we may consider cases where σ scales with N . We then denote the coefficient σ_N and assume that $\sigma_N \rightarrow \sigma \geq 0$.

Observe that the Wiener processes are only present in the velocity equations which will have several important consequences.

The space domain Γ may be the whole space \mathbb{R}^d , the flat torus \mathbb{T}^d , or some bounded domain. The analysis of a bounded, smooth domain Γ is strongly dependent on the

type of boundary conditions but can sometimes be handled in a manner similar to the other cases with some adjustments. Thus, for simplicity, we typically limit ourselves to $\Gamma = \mathbb{R}^d$, \mathbb{T}^d . Even if Γ is bounded, there is no hard cap on velocities so that the actual domain in position and velocity, $\Gamma \times \mathbb{R}^d$, is always unbounded.

The critical scaling in (1) (and later in (3)) is the factor $\frac{1}{N}$ in front of the interaction terms. This is *the mean field scaling*, and it keeps, at least formally, the total strength of the interaction of order 1.

At least formally, one expects that as the number N of particles goes to infinity, (1) will be replaced by a continuous PDE. In the present case, the candidate is the so-called McKean–Vlasov equation (or sometimes Vlasov–Fokker–Planck) which reads

$$\partial_t f + v \cdot \nabla_x f + (K \star \rho) \cdot \nabla_v f = \sigma \Delta_v f, \quad (2)$$

where the unknown $f = f(t, x, v)$ is the phase space density or one-particle distribution and $\rho = \rho(t, x)$ is the spatial (macroscopic) density obtained through

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

This type convergence is what we call *mean field limit*, and it is connected to the important property of *propagation of chaos*.

We point out that Eq. (2) is of degenerate parabolic type as the diffusion $\Delta_v f$ only acts on the velocity variable.

1.2 First-Order Systems

As the companion of (1), we also consider the first-order stochastic system

$$dX_i = \frac{1}{N} \sum_{j \neq i} K(X_i - X_j) dt + \sqrt{2\sigma} dW_t^i, \quad i = 1, \dots, N, \quad (3)$$

with the same assumptions as for the system (1). As before, one expects that as the number N of particles goes to infinity, the system (3) will converge to the following PDE

$$\partial_t f + \operatorname{div}_x(f(K \star f)) = \sigma \Delta_x f, \quad (4)$$

where the unknown $f = f(t, x)$ is now the spatial density.

The model (3) can be regarded as the small mass limit (Smoluchowski–Kramers approximation) of Langevin equations in statistical physics. However, the model (3) has its own important applications.

The best-known classical application is in fluid dynamics with the Biot–Savart kernel

$$K(x) = \frac{1}{2\pi} \left(\frac{-x_2}{|x|^2}, \frac{x_1}{|x|^2} \right).$$

This leads to the well-known vortex model which is widely used to approximate the 2D Navier–Stokes equation written in vorticity form. See for instance [14, 15, 29, 59, 67].

System (1) or (3) can be written in the more general form of

$$dZ_i = \frac{1}{N} \sum_{j=1}^N H(Z_i, Z_j) dt + \sqrt{2\sigma} d\tilde{W}_t^i, \quad i = 1, 2, \dots, N, \quad (5)$$

where we denote $\tilde{W}_t^i = W_t^i$ in the case of first-order models and $\tilde{W}_t^i = (0, W_t^i)$ for second-order models where there is only diffusion in velocity.

It is easy to check that taking

$$H(Z_i, Z_j) = K(Z_i - Z_j)$$

with the convention that $K(0) = 0$ and the system of SDEs (5) becomes (3). Furthermore, if we replace Z_i by the pair (X_i, V_i) as in (1) and set H as

$$H((X_i, V_i), (X_j, V_j)) = (V_i, K(X_i - X_j))$$

and again with convention that $K(0) = 0$, then the SDEs (5) can also represent the Newton-like systems (1). We write particle systems in the forms as (1) and (3) simply because that is enough for most interesting applications.

1.3 Examples of Applications

As mentioned above, the best-known example of interaction kernel is the Poisson kernel, that is,

$$K(x) = \pm C_d \frac{x}{|x|^d}, \quad d = 2, 3, \dots,$$

where $C_d > 0$ is a constant depending on the dimension and the physical parameters of the particles (mass, charges, etc.). This corresponds to particles under gravitational interactions for the case with a minus sign and electrostatic interactions (ions in a plasma for instance) for the case with a positive sign.

The description by McKean–Vlasov PDEs goes far beyond plasma physics and astrophysics. Large systems of interacting particles are now widely used in the bio-sciences and social sciences, since many individual-based particle models are formulated to model collective behaviors of self-organizing particles or agents. Given

the considerable literature, we only give a few limited examples and the references that are cited have no pretension to be exhaustive.

- Biologic systems model the collective motion of microorganisms (bacteria, cells, etc.). The canonical example is again the case of the Poisson kernel for K where system (3) coincides with the particle models to approximate the Keller–Segel equation of chemotaxis. We refer mainly to [30] for the mean field limit, together with [34, 56] (see also [68, 69, 72] for general modeling discussions).
- Aggregation models correspond to system (3), typically with $K = -\nabla W$ and an extra potential term $-\nabla V(X_i)$,

$$dX_i = -\frac{1}{N} \sum_{j \neq i} \nabla W(X_i - X_j) dt - \nabla V(X_i) dt + \sqrt{2\sigma} dW_t^i, \quad i = 1, \dots, N. \quad (6)$$

They are used in many settings (in biology, ecology, for space homogeneous granular media as in [3], etc.). See for instance [9, 10, 13, 57, 58] for the mathematical analysis of the particle system (6) [19, 20, 25], for the analysis of the limiting PDE.

- Since the pioneering works in [74] and [24], second-order systems like (1) have been used to model flocks of birds, schools of fishes, swarms of insects, etc. One can see [17, 18, 40] and the references therein for a more detailed discussion of flocking or swarming models in the literature. We emphasize that the presence of noise in the models is important since we cannot expect animals to interact with each other or the environment in a completely deterministic way. We in particular refer to [39] for stochastic Cucker–Smale model with additive white noise as in (1) and to [1] for multiplicative white noise in velocity variables, respectively. The rigorous proof of the mean field limit was given in [7] for systems similar to (1) with locally Lipschitz vector fields; the mean field limit for stochastic Vicsek model where the speed is fixed is given in [8].
- First-order models are quite popular to model opinion dynamics among a population (such as the emergence of a common belief in a pricing system). We refer for instance to [43, 52, 65, 77]. Individual-based models are even used for coordination or consensus algorithm in control engineering for robots and unmanned vehicles (see [21]).

There are many other interesting questions that are related to mean field limit for stochastic systems but that are out of the scope of the present article. For instance,

- *The derivation of collisional models and Kac's program in kinetic theory.* After the seminal in [53] and later [23], the rigorous derivation of the Boltzmann equation was finally achieved in [31] but only for short times (of the order of the average time between collisions). The derivation for longer times is still widely open in spite of some critical progress when close to equilibrium in [4, 5]. Many tools and concepts that are used for mean field limits were initially introduced for collisional models, such as the ideas in the now famous Kac's program. Kac first introduced a probabilistic approach to simulate the spatially homogeneous Boltzmann equation

in [50] and formulated several related conjectures. After some major contributions in the 1990s, see in particular [26, 37, 38], significant progress was again achieved recently in solving these conjectures, see [41, 63, 64], and the earlier [16].

- Stochastic vortex dynamics with multiplicative (instead of additive) noise leading to stochastic 2D Euler equation. In [28], the authors showed that the point vortex dynamics become fully well-posed for every initial configuration when a generic stochastic perturbation (in the form of multiplicative noises) compatible with the Euler description is introduced. The SDE systems in [28] will converge to the stochastic Euler equation, rather than Navier–Stokes equation as the number N of point vortices goes to infinity. However, the rigorous proof of the convergence is difficult and still open.
- Scaling limit (hydrodynamic limit) of random walks on discrete spaces, for instance on lattice \mathbb{Z}^d for which we refer to [51]. In this setting, one also tries to obtain a continuum model, usually a deterministic PDE, from a discrete particle model on a lattice, as $N \rightarrow \infty$ and of course the mesh size h converges to 0. An interesting observation is that we can use a stochastic PDE as a correction to the limit deterministic PDE (see [27]).

This article is organized as follows: In Section 2, we introduce the basic concepts and tools in this subject. We define the classical notion of mean field limit and propagation of chaos as well as more recent notions of chaos. Then, in Section 3, we review some classical results under the assumptions that K is globally Lipschitz and more recent results for singular kernels K . Both qualitative and quantitative results will be presented. Finally, in Section 4, we briefly review the authors’ recent results [47] and [48] for very rough interaction kernels K with the relative entropy method.

2 The Basic Concepts and Main Tools

In order to compare the particle systems (1) and (3) with the expected mean field equations (2) and (4), respectively, we need to introduce several concepts and tools to capture the information on both levels of descriptions.

Those tools have often been introduced in different contexts (and in particular for the derivation of collisional models such as the Boltzmann equation). The main classical references here are [6, 36, 50, 71] and [45] for the stochastic aspects.

2.1 *The Empirical Measure*

In the following, to make the presentation simple, we sometimes use the unified, one-variable formulation (5). Therefore, for the second-order model,

$$Z_i = (X_i, V_i) \in E := \Gamma \times \mathbb{R}^d$$

and for the first-order model,

$$Z_i = X_i \in E := \Gamma.$$

One defines the empirical measure as

$$\mu_N(t, z) = \frac{1}{N} \sum_{i=1}^N \delta(z - Z_i(t)), \quad (7)$$

where $z = (x, v) \in E = \Gamma \times \mathbb{R}^d$ or $z = x \in \Gamma$.

The empirical measure is a random probability measure, *i.e.*, $\mu_N(t) \in \mathcal{P}(E)$, whose law lies in the space $\mathcal{P}(\mathcal{P}(E))$. Recall that $\mathcal{P}(E)$ represents the set of all Borel probability measures on E . Since all particles Z_i are assumed to be indistinguishable, $\mu_N(t, z)$ gives the full information of the solution

$$Z^N(t) = (Z_1(t), \dots, Z_N(t)),$$

to the particle system (1) or (3).

One uses a slight variant in the case of 2D stochastic vortex model for approximating Navier–Stokes equation where for convenience one usually defines

$$\mu_N(t, x) = \frac{1}{N} \sum_{i=1}^N \alpha_i \delta(x - X_i(t)) \quad (8)$$

(see for instance [29]). In that case, α_i models the strength of the circulation which can be positive or negative, and hence, the empirical measure is a signed measure and not a probability measure.

In the deterministic setting, that is provided $\sigma = 0$ in (1) or (3) (and therefore in (2) and (4)), the empirical measure μ_N is also deterministic. Furthermore, in this special case, it actually solves exactly the limiting PDE (2) or (4) in the sense of distribution. However, this cannot be true anymore for the stochastic setting: The stochastic behavior can only vanish when the number N of particles goes to infinity.

Systems (1) and (3) need to be supplemented with initial conditions. A first possibility is to choose a deterministic sequence of initial data $Z^N(t = 0)$.

It is considered more realistic though to use random initial conditions in which case $Z^N(t = 0)$ is taken according to a certain law

$$\text{Law}(Z_1^0, \dots, Z_N^0) = F^N(0) \in \mathcal{P}(E^N).$$

One often assumes some sort of independence (or almost independence) in the law $F^N(0)$. A typical example is *chaotic law* for which $F^N(0, z_1, \dots, z_N) = \prod_{i=1}^N f^0(z_i)$ for a given function f^0 .

No matter which type of initial condition is chosen, $\mu_N(t, z)$ is always random for any $t > 0$.

The concept of mean field limit is defined for a particular choice of sequence of initial data.

Definition 1 (Mean field limit) Consider a sequence of deterministic initial data (Z_1^0, \dots, Z_N^0) as $N \rightarrow \infty$ or equivalently a sequence of deterministic initial empirical measures μ_N^0 , such that

$$\mu_N^0 \rightarrow f_0$$

in the tight topology of measures. Then, the mean field limit holds for this particular sequence iff for *a.e.* t

$$\mu_N(t) \rightarrow f_t$$

with convergence in law, where f_t denotes the solution to (2) or (4) with initial data f_0 .

Here, the convergence in law of μ_N means that for any ϕ bounded continuous function from the set of probability measures on E , $\mathcal{P}(E)$ to \mathbb{R} , one has that

$$\phi(\mu_N(t, .)) \longrightarrow \phi(f_t(.)), \quad (9)$$

as $N \rightarrow \infty$, where $t \geq 0$. This in particular implies the convergence of the first moment (the case where ϕ is linear in μ_N) s.t. for any $\varphi \in C_b(E)$

$$\mathbb{E} \int_E \varphi(z) d\mu_N(t, z) := \mathbb{E} \frac{1}{N} \sum_{i=1}^N \varphi(Z_i(t)) \rightarrow \int_E \varphi(z) f_t(z) dz, \quad (10)$$

A few important points follow from this definition:

- The mean field limit may hold for one particular choice of sequence and not hold for another. In fact, for many singular kernels, this is very likely as it is easy to build counterexamples where the particles are initially already concentrated. This means that the right questions are *for which sequence of initial data the mean field limit holds* and whether *this set of initial data are somehow generic*.
- If one chooses random initial data for instance according to some chaotic law $F^N(0) = (f^0)^{\otimes N}$, then the question of the mean field limit can be asked for each instance of the initial data. Hence, the mean field limit could *a priori* hold with a certain probability that one would hope to prove equal to 1. This will lead to the important notion of propagation of chaos.
- Because the limit f is deterministic (as a solution to a PDE), the mean field limit obviously only holds if μ_N becomes deterministic at the limit. This hints at possible connection with some sort of law of large numbers.

2.2 The Liouville Equation

While the empirical measure follows the trajectories of the system, it can be useful to have statistical information as given by the joint law

$$F^N(t, z_1, \dots, z_N) = \text{Law}(Z_1(t), \dots, Z_N(t))$$

of the particle systems (1) or (3). F^N is not experimentally measurable for practical purposes, and instead, the observable statistical information of the system is contained in its marginals. One hence defines k -marginal distribution as

$$F_k^N(t, z_1, \dots, z_k) = \int_{E^{N-k}} F^N(t, z_1, \dots, z_N) dz_{k+1} \dots dz_N. \quad (11)$$

The 1-marginal is also known as the one-particle distribution, while the 2-marginal contains the information about pairwise correlations between particles.

It is possible to write a closed equation, usually called the Liouville equation, governing the evolution of the law F^N . For second-order systems, it reads

$$\partial_t F^N + \sum_{i=1}^N v_i \cdot \nabla_{x_i} F^N + \frac{1}{N} \sum_{i=1}^N \sum_{j \neq i} K(x_i - x_j) \cdot \nabla_{v_i} F^N = \sigma \sum_{i=1}^N \Delta_{v_i} F^N. \quad (12)$$

Similarly, for the first-order systems (3), one has

$$\partial_t F^N + \frac{1}{N} \sum_{i=1}^N \sum_{j \neq i} K(x_i - x_j) \cdot \nabla_{x_i} F^N = \sigma \sum_{i=1}^N \Delta_{x_i} F^N. \quad (13)$$

In the deterministic case, those equations had been derived by Gibbs (see [32, 33]). In the present stochastic setting, they follow from Itô's formula, see [45], applied to $\phi(Z^N(t))$ for any test function.

The fact that particles are indistinguishable implies that F^N is a symmetric probability measure on the space E^N , that is, for any permutation of indices $\tau \in S_N$,

$$F^N(t, z_1, \dots, z_N) = F^N(t, z_{\tau(1)}, \dots, z_{\tau(N)}).$$

We write it as $F^N \in \mathcal{P}_{\text{Sym}}(E^N)$. Similarly, it is easy to check that the k -marginal distribution is also symmetric $F_k^N \in \mathcal{P}_{\text{Sym}}(E^k)$ for $2 \leq k \leq N$.

2.3 The BBGKY Hierarchy

For simplicity, we only focus on the first-order system (1) as the second-order dynamics (1) case can be dealt with similarly by adding the corresponding free transport terms.

From the Liouville equation (13), it is possible to deduce equations on each marginal F_k^N . Noticing the fact that particles are indistinguishable, and using the appropriate permutation, one obtains

$$\begin{aligned} \partial_t F_k^N + \frac{1}{N} \sum_{i=1}^k \sum_{j=1, j \neq i}^k K(X_i - X_j) \cdot \nabla_{x_i} F_k^N \\ + \frac{N-k}{N} \sum_{i=1}^k \int_{\Gamma} \operatorname{div}_{x_i} \left(K(x_i - y) F_{k+1}^N(t, x_1, \dots, x_k, y) \right) dy = \sigma \sum_{i=1}^k \Delta_{x_i} F_k^N. \end{aligned} \quad (14)$$

The equation (14) is not closed as it involves the next marginal F_{k+1}^N and thus the denomination of hierarchy.

On the other hand, each marginal F_k^N is defined on a fixed space E^k contrary to F^N which is defined on a space depending on N . Therefore, one may easily consider the limit of F_k^N as $N \rightarrow \infty$ for a fixed k . Formally, one obtains the limiting hierarchy

$$\partial_t F_k^\infty + \sum_{i=1}^k \int_{\Gamma} \operatorname{div}_{x_i} \left(K(x_i - y) F_{k+1}^\infty(t, x_1, \dots, x_k, y) \right) dy = \sigma \sum_{i=1}^k \Delta_{x_i} F_k^\infty.$$

Each equation is still not closed, and *a priori*, the hierarchy would have to be considered for all k up to infinity.

However, if $F_k^\infty(t)$ is tensorized, $F_k^\infty(t) = f_t^{\otimes k}$, then each equation is closed and they all reduce to the limiting mean field equation (4). This leads us to the important notion of propagation of chaos.

2.4 Various Notions of Chaos

The original notion of propagation of chaos goes as far back as Maxwell and Boltzmann. The classical notion of propagation of chaos was formalized by Kac in [50] (see also the famous [71]). The other, stronger notions of chaos presented here were investigated more recently in particular in [41], [63], and [64] (see also [16]).

Let us begin with the simplest definition that we already saw.

Definition 2 A law F^N is tensorized/factorized/chaotic if

$$F^N(z_1, \dots, z_N) = \prod_{i=1}^N F_1^N(z_i).$$

Unfortunately, for N fixed, the law $F^N(t)$ solving the Liouville Eq. (12) or (13) cannot be chaotic. Indeed, for a fixed N , some measure of dependence necessarily exists between particles and strict independence is only possible asymptotically. This leads to Kac's chaos.

Definition 3 Let E be a measurable metric space (here $E = \Gamma \times \mathbb{R}^d$ or Γ). A sequence $(F^N)_{N \in \mathbb{N}}$ of symmetric probability measures on E^N is said to be f -chaotic for a probability measure f on E , if one of the following equivalent properties holds:

- i) For any $k \in \mathbb{N}$, the k -marginal F_k^N of F^N converges weakly toward $f^{\otimes k}$ as N goes to infinity, i.e., $F_k^N \rightharpoonup f^{\otimes k}$;
- ii) The second marginal F_2^N converges weakly toward $f^{\otimes 2}$ as N goes to infinity: $F_2^N \rightharpoonup f^{\otimes 2}$;
- iii) The empirical measure associated with F^N , that is $\mu_N(z) \in \mathcal{P}(E)$ as in (7) with $F^N = \text{Law}(Z_1, \dots, Z_N)$, converges in law to the deterministic measure f as N goes to infinity.

Here, the weak convergence $F_k^N \rightharpoonup f^{\otimes k}$ simply means that for any test functions $\phi_1, \dots, \phi_k \in C_b(E)$,

$$\lim_{N \rightarrow 0} \int_{E^k} \phi_1(z_1) \cdots \phi_k(z_k) F_k^N(z_1, \dots, z_k) dz_1 \cdots dz_k = \prod_{i=1}^k \int_E f(z_i) \phi_i(z_i) dz_i,$$

and μ_N converges in law to f as before in the sense of (9).

We refer to [71] for the classical proof of equivalence between the three properties. A version of the equivalence has recently been obtained in [41], quantified by the 1 Monge–Kantorovich–Wasserstein (MKW) distance between the laws.

We now can define the corresponding notion of propagation of chaos.

Definition 4 (Propagation of chaos) Assume that the sequence of the initial joint distribution $(F^N(0))_{N \geq 2}$ is f_0 -chaotic. Then, propagation of chaos holds for systems (1) or (3) up to time $T > 0$ iff for any $t \in [0, T]$, and the sequence of the joint distribution at time t $(F^N(t))_{N \geq 2}$ is also f_t -chaotic, where f_t is the solution to the mean field PDE (2) or (4), respectively, with initial data f_0 .

If initially (Z_1^0, \dots, Z_N^0) was chosen according to the law $F^N(0)$ with the sequence $F^N(0)$ to be f_0 -chaotic, then by property iii) of Definition 3 propagation of chaos implies that the mean field limit holds with probability one.

Kac's chaos and propagation of chaos are rather weak and thus do not allow a very precise control on the initial data. For this reason, it can be useful to consider stronger notions of chaos.

There are two natural physical quantities that can help quantify such stronger notions of chaos: the (Boltzmann) entropy and the Fisher information. The scaled entropy of the law F^N is defined as

$$H_N(F^N) = \frac{1}{N} \int_{E^N} F^N \log F^N dz_1 \cdots dz_N,$$

where we recall that $E = \Gamma \times \mathbb{R}^d$ for the second-order system and $E = \Gamma$ for the first-order system. The Fisher information is

$$I_N(F^N) = \frac{1}{N} \int_{E^N} \frac{|\nabla F^N|^2}{F^N} dz_1 \cdots dz_N.$$

We normalized both quantities by factor $\frac{1}{N}$ such that for any $f \in \mathcal{P}(E)$,

$$H_N(f^{\otimes N}) = H_1(f), \quad I_N(f^{\otimes N}) = I_1(f).$$

The use of those quantities leads to alternative and stronger definitions of a f -chaotic sequence, *entropy chaos*, and *Fisher information chaos*.

Definition 5 (Definition 1.3 in [41]) Consider $f \in \mathcal{P}(E)$ and a sequence $(F^N)_{N \geq 2}$ of $\mathcal{P}_{\text{Sym}}(E^N)$ such that for some $k > 0$, the k th moment $M_k(F_1^N) = \int |z|^k dF_1^N$ of F_1^N is uniformly bounded in N . We say that

i) the sequence (F^N) is f -Fisher information chaotic if

$$F_1^N \rightharpoonup f, \quad I_N(F^N) \rightarrow I_1(f), \quad I_1(f) < \infty$$

ii) the sequence (F^N) is f -entropy chaotic if

$$F_1^N \rightharpoonup f, \quad H_N(F^N) \rightarrow H_1(f), \quad H_1(f) < \infty.$$

There are even intermediary notions that we omit for simplicity. There exists a strict hierarchy between these two definitions as per

Theorem 1 (Theorem 1.4 in [41]) Consider $f \in \mathcal{P}(E)$ and $(F^N)_{N \geq 2}$ a sequence of $\mathcal{P}_{\text{Sym}}(E^N)$ such that the k th moment $M_k(F_1^N)$ is bounded, for some $k > 2$. In the list of assertions below, each one implies the assertion which follows:

- i) (F^N) is f -Fisher information chaotic;
- ii) (F^N) is f -Kac's chaotic (as in Definition 3) and $I_N(F^N)$ is bounded;
- iii) (F^N) is f -entropy chaotic; and
- iv) (F^N) is f -Kac's chaotic.

For some recent results on propagation of chaos in strong sense, we refer to [29] where the convergence is in the sense of entropy as in ii) in Definition 5 and to [34] for a similar argument for the subcritical Keller–Segel model.

We will finish this section by considering the relation between the entropy of the full joint law and the entropy of the marginals.

Proposition 1 For any $F^N \in \mathcal{P}_{\text{Sym}}(E^N)$ and $f \in \mathcal{P}(E)$, one has

$$H_k(F_k^N) \leq H_N(F^N), \quad H_k(F_k^N | f^{\otimes k}) \leq H_N(F^N | f^{\otimes N}). \quad (15)$$

The scaled relative entropy is defined by

$$H_N(F^N | f^{\otimes N}) = \frac{1}{N} \int_{E^N} F^N \log \frac{F^N}{f^{\otimes N}}.$$

By Proposition 1 and the Liouville equation (12) or (13), we can control the entropy of any marginal at any time $t > 0$ given the uniform bound $\sup_{N \geq 2} H_N(F^N(0))$ initially. This is really surprising since for instance the free bound

$$\sup_{N \geq 2} H_k(F_k^N(t, \cdot)) < \infty$$

will *a priori* ensure that the weak limit f_1 of F_1^N belongs to $L^1(E)$ at any time $t > 0$ by Theorem 3.4 in [41], without having to know or prove anything about the mean field limit.

3 Some of the Main Results on Mean Field Limits

In this section, we review some main results on mean field limits for stochastic particle systems. The classical results, such as the famous [60] or [71] (see also the very nice presentation in [61]), will normally require that the interaction kernel K be Lipschitz; we also refer to [70]. For singular (not locally Lipschitz) kernels, only a few results are available, mostly in the context of 2D stochastic vortex model (see for instance [29, 59, 62, 67]). We also refer to [14, 15, 22] and [30, 34, 56] for the singular kernel cases.

3.1 The Classical Approach: Control on the Trajectories

The results we present here are taken mainly from [71] and [61]. Note though that the diffusion processes considered in [61] are much more general than what we use here

$$dX_i = \frac{1}{N} \sum_{i=1}^N b(X_i, X_j) dt + \frac{1}{N} \sum_{i=1}^N \sigma(X_i, X_j) dW_t^i$$

where $i = 1, 2, \dots, N$, and the vector field b and matrix field σ are Lipschitz continuous with respect to both variables. As before, the W_t^i are mutually independent d -dimensional Brownian motions.

However, to make the presentation simple, we only focus on the first-order system (3) assuming that the kernel K is *Lipschitz*. Second-order systems with the same Lipschitz assumption can be treated in a similar manner.

If the interaction kernel K is globally Lipschitz, a standard method to show the mean field limit was popularized by Sznitman [71] (see also the more recent [58]). We also refer to [7, 9, 12] and the reference therein for recent developments and to [64] for the quantitative Grunbaum's duality method.

The basic idea of the method is as follows: For system (3) endowed with initial data

$$X_i(0) = X_i^0, \quad i.i.d. \text{ with Law } X_i^0 = f_0,$$

we construct a symmetric particle system coupled to (3), that is,

$$d\bar{X}_i = K \star f_t(\bar{X}_i) + \sqrt{2\sigma} dW_t^i, \quad \bar{X}_i(0) = X_i^0, \quad i = 1, 2, \dots, N, \quad (16)$$

where the W_t^i are the same Brownian motions as in (3) and $f_t = \text{Law}(\bar{X}_i(t))$. The coupling between (3) and (16) is only due to the fact that they have the same initial data and share the same Brownian motions.

Observe that Eq. (16) is not anymore an SDE system: The dynamics between particles are now coupled through the law f_t . That law is the same for each \bar{X}_i , and in that sense, one only considers N independent copies of the same system given by

$$dX_t = K \star f_t(X_t) + \sqrt{2\sigma} dW_t, \quad \text{Law}(X_t) = f_t. \quad (17)$$

It is straightforward to check that f_t is just the (weak) solution to the mean field PDE (4) by Itô's formula. Therefore, the law of large numbers in the right function space will give us that the system (16) is close to the mean field PDE (4).

Under some stability estimates that follow the traditional Gronwall type of bounds for SDEs, it can be shown that the symmetric system (16) is close to the original one (3).

On the other hand, (17) and (4) are well-posed with existence and uniqueness, which finally implies the mean field limit. We summarize the results of the above discussion with the following two theorems.

Theorem 2 (Theorem 1.1 in [71] and Theorem 2.2 in [61]) *Assume that K is globally Lipschitz and f_0 is a Borel probability measure on \mathbb{R}^d with finite second moment. Then, there is existence and uniqueness of the solutions to (17) as well as to (4).*

We refer to [71] for a detailed proof in the case where K is also bounded and to [61] for a proof in the general case. We remark that the existence and uniqueness hold both trajectory-wise and in law for (17).

With Theorem 2 in mind, as long as we can show that the systems (3) and (16) are close, we will obtain the mean field limit. This is provided by the following theorem.

Theorem 3 (Theorem 2.3 in [61] and Theorem 1.4 in [71]) *Assume that K is globally Lipschitz and f_0 is a Borel probability measure on \mathbb{R}^d with finite second moment. Then, for any $i = 1, \dots, N$, one has*

$$\mathbb{E} \left(\sup_{0 \leq t \leq T} |X_i(t) - \bar{X}_i(t)|^2 \right) \leq \frac{C}{N}, \quad (18)$$

where C is independent of N , but depends on the time interval T and the Lipschitz constant $\|K\|_{Lip}$.

We omit the proof and instead remark that Theorem 2 obviously implies the mean field limit or propagation of chaos. Recall that the p -MKW distance between two probability measures μ and ν with finite p th moments is defined by

$$W_p(\mu, \nu) = \inf_{(X, Y)} (\mathbb{E}|X - Y|^p)^{\frac{1}{p}},$$

where the infimum runs over all couples of random variables (X, Y) with $\text{Law } X = \mu$ and $\text{Law } Y = \nu$ (see for instance [73]).

From Theorem 3, one obtains an estimate on the distance between the 1-marginal F_1^N and f_t as $N \rightarrow \infty$,

$$W_2^2(F_1^N, f_t) \leq \mathbb{E}|X_i(t) - \bar{X}_i(t)|^2 \leq \frac{C}{N}.$$

More generally, one has a quantitative version of propagation of chaos. For any fixed k , the k -marginal distribution F_k^N converges to $f^{\otimes k}$ as N goes to infinity

$$W_2^2(F_k^N(t), (f_t)^{\otimes k}) \leq \mathbb{E} |(X_1(t), \dots, X_k(t)) - (\bar{X}_1(t), \dots, \bar{X}_k(t))|^2 \leq \frac{kC}{N}.$$

Similarly, we can obtain the convergence in law of the empirical measure toward the limit f_t . Indeed, for a test function $\phi \in C_b^1(\mathbb{R}^d)$, one has

$$\begin{aligned} & \mathbb{E} \left| \frac{1}{N} \sum_{i=1}^N \phi(X_i(t)) - \int_{\mathbb{R}^d} \phi(x) f_t(x) dx \right|^2 \\ & \leq 2\mathbb{E}|\phi(X_1(t)) - \phi(\bar{X}_1(t))|^2 + 2\mathbb{E} \left| \frac{1}{N} \sum_{i=1}^N \phi(\bar{X}_i(t)) - \int_{\mathbb{R}^d} \phi(x) f_t(x) dx \right|^2 \leq \frac{C}{N} \|\phi\|_{C^1}. \end{aligned} \tag{19}$$

3.2 Large Deviation and Time Uniform Estimates

The results referred to here were mostly obtained in [9, 10, 13, 57, 58]. The average estimates for instance (18) and (19) guarantee that the particle system (1) or (3) is an approximation to the mean field PDE (2) or (4), respectively, for a fixed time interval.

However, uniformity in time estimates is sometimes necessary: Make sure that the equilibrium for the discrete system is close to the equilibrium of the continuous model for instance.

Furthermore, it can be critical to be able to estimate precisely how likely any given instance of the discrete system is to be far from the limit. When large stochastic particle systems are used to do numerical simulations, one may want to make sure that the numerical method has a very small probability to give wrong results. Those are usually called concentration estimates.

Unfortunately, the bound (18) can only give very weak concentration estimates by Chebyshev's inequality, for instance

$$\mathbb{P} \left\{ |X_i(t) - \bar{X}_i(t)| \geq \frac{\sqrt{C} L}{\sqrt{N}} \right\} \leq \frac{1}{L^2}. \quad (20)$$

Uniformity in time estimates in particular cannot hold for any system. For this reason, we consider here a particular variant of first-order particle system (3), namely system (6) with $\sigma = 1$ and with V and W convex. In that case, the mean field equation corresponding to system (6) is

$$\partial_t f = \Delta_x f + \operatorname{div}_x(f \nabla_x(V + W \star f)). \quad (21)$$

One similarly constructs a symmetric coupling system

$$d\bar{X}_i = -\nabla W \star f_t(\bar{X}_i) - \nabla V(\bar{X}_i) + \sqrt{2} dW_t^i, \quad \bar{X}_i(0) = X_i^0, \quad (22)$$

where $f_t = \operatorname{Law}(\bar{X}_i(t))$ and is hence the solution to (21) with initial data f_0 . Then, one can obtain the following theorem.

Theorem 4 (Theorem 1.2, Theorem 1.3, and Proposition 3.22 in [57]) Assume that the interaction potential W is convex, even, and with polynomial growth and V is uniformly convex i.e., $D^2V(x) \geq \beta I$ for some $\beta > 0$. Assume in addition that initially,

$$X_i(0) = X_i^0, \quad i.i.d., \quad \operatorname{Law}X_i^0 = f_0,$$

where f_0 is smooth. Then, there exists a constant C such that for any $N \geq 2$,

$$\sup_{t>0} \mathbb{E}(|X_i(t) - \bar{X}_i(t)|^2) \leq \frac{C}{N}, \quad (23)$$

and for any $\varepsilon > 0$,

$$\sup_{\|\phi\|_{Lip} \leq 1} \mathbb{P} \left[\left| \frac{1}{N} \sum_{i=1}^N \phi(X_i(t)) - \int \phi(x) f_t(x) dx \right| > \frac{C}{\sqrt{N}} + \varepsilon \right] \leq 2 \exp(-\frac{\beta}{2} N \varepsilon^2). \quad (24)$$

Compare the exponential control on the tail in (24) to the polynomial estimate in (20). This large deviation-type estimate (24) for the empirical measure is obtained by the use of logarithmic Sobolev inequalities.

Under the above convexity assumptions on the potentials V and W , the solution f_t to the granular media equation (21) converges to a unique equilibrium exponentially fast. It is in this context that a uniform in time estimate can be expected.

In [10], a stronger version of (24) was achieved, but at the same time, more restrictions were imposed on V and W and the initial law f_0 . At least if W and V do not grow too fast, the following theorem holds.

Theorem 5 (Theorem 2.9 in [10]) *Assume that V and W both are uniformly convex and have appropriate growth at infinity. Assume that initially,*

$$X_i(0) = X_i^0, \quad i.i.d., \quad \text{Law} X_i^0 = f_0,$$

for a smooth f_0 with a finite square exponential moment, i.e., there exists $\alpha_0 > 0$, such that

$$\int \exp(\alpha_0|x|^2) f_0(x) dx < \infty.$$

Then for any $T > 0$, there exists a constant $C = C(T)$ such that for any $d' > d$, there exist some constants N_0 and C' such that for all $\varepsilon > 0$, if $N \geq N_0 \max(\varepsilon^{-(d'+2)}, 1)$, then

$$\mathbb{P} \left[\sup_{0 \leq t \leq T} W_1(\mu_N(t), f_t) > \varepsilon \right] \leq C'(1 + T\varepsilon^{-2}) \exp(-C N \varepsilon^2). \quad (25)$$

In the above theorem, W_1 denotes the 1 MKW distance. While the constants are now time dependent, the result is more precise. It is even possible to estimate the deviation on the empirical measure on pairs of particles, so that

$$\mu_N^2(t) = \frac{1}{N(N-1)} \sum_{i \neq j} \delta_{(X_i(t), X_j(t))}$$

is close to $f_t^{\otimes 2}$ in the sense of (25). See Theorem 2.10 in [10] and also Theorem 2.12 there for uniformity in time estimates in the spirit of (25).

3.3 Singular Kernels: Stochastic Vortex Model Leading to 2D Navier–Stokes Equation

In this subsection, we review some results on the mean field limit for stochastic systems with singular kernels K : K is smooth on $\mathbb{R}^d \setminus \{0\}$, but in general, $|K(x)| \rightarrow \infty$ when $x \rightarrow 0$. Therefore, when two particles are close, the interaction between them becomes extremely large.

As mentioned in the introduction, the second-order systems are generally harder, as diffusion is now degenerate. For the Poisson kernel and particles interacting under gravitational force or Coulomb force, both the deterministic case (leading to Vlasov equation) and the stochastic case (leading to McKean–Vlasov equation) are still open.

However, the first-order systems are usually easier. There are several mean field limit results for first-order systems with

$$K(x) \sim \frac{1}{|x|^\alpha}, \quad \text{when } x \sim 0$$

for some $\alpha > 0$. We refer for instance [30, 34] for Keller–Segel equation and [56] for random particle blob method also in the setting of Keller–Segel equation. For other singular kernels, one can also see [22] about the Coulomb gas model in 1D, where the singularity is repulsive and strong, behaving like $\frac{1}{|x|}$.

We focus in the rest of this subsection on the Biot–Savart kernel K which leads to the vortex model in fluid mechanics in dimension 2. This case is now better understood, thanks for instance to [59, 62, 67] and more recently [29]. The results presented here are mainly based on [67] and [29].

In general, the stochastic or viscous vortex model is written as

$$dX_i = \frac{1}{N} \sum_{j \neq i} \alpha_j K(X_i - X_j) dt + \sqrt{2\sigma} dW_t^i, \quad (26)$$

and the empirical measure is defined as in (8) to more easily allow for a negative vorticity. However, for simplicity, we here assume that all $\alpha_i \equiv 1$, and hence, (26) reduces to our classical first-order system (3) with

$$K(x) = \frac{1}{2\pi} \left(\frac{-x_2}{|x|^2}, \frac{x_1}{|x|^2} \right).$$

The expected mean field PDE given by (4) is now the 2D Navier–Stokes equation in vorticity formulation with positive viscosity σ .

It is not initially obvious that there even exists solutions for a fixed N because of the singularity in K . In [66] (see also Theorem 2.10 in [29]), it was showed that almost surely for all $t \geq 0$, and all $i \neq j$, $X_i(t) \neq X_j(t)$. Hence, the system (26) is well-posed since the singularity of K is almost surely never visited.

The main result from [67] is the propagation of chaos

Theorem 6 *Assume that $(F^N(0))_{N \geq 2}$ is f_0 -chaotic and*

$$\limsup_{N \rightarrow \infty} \left\| \int_{(\mathbb{R}^2)^{N-i}} F^N(0) dx_{i+1} \cdots dx_N \right\|_{L^\infty((\mathbb{R})^i)} < \infty, \quad \text{for } i = 1, 2. \quad (27)$$

Then, there exists $\sigma_0 > 0$ such that $(F^N(t))_{N \geq 2}$ is f_t -chaotic for any $\sigma > \sigma_0$. In an equivalent way, we have $\mu_N(t) \rightarrow f(t)$ in law as $N \rightarrow \infty$ for $\sigma > \sigma_0$.

More recently, the above result was improved in [29]: No assumption is required on the viscosity and the initial vorticity f_0 belongs to $L^1(\mathbb{R}^2)$, while in [67] it is essentially required that $f_0 \in L^\infty$. We state a simplified version of the main theorem from [29].

Theorem 7 (Theorem 2.12 and Theorem 2.13 in [29]) Consider any $f_0 \geq 0$, an initial data for (4) with

$$\int_{\mathbb{R}^2} f_0 (1 + |x|^k + |\log f_0|) dx < \infty, \quad \text{for some } k \in (0, 2).$$

Assume that for $N \geq 2$, the law $F^N(0)$ of the initial distribution of particles is f_0 -chaotic and

$$\begin{aligned} \sup_{N \geq 2} \frac{1}{N} \int_{(\mathbb{R}^2)^N} (1 + |X|^2)^{\frac{k}{2}} F^N(0, X) dX &< \infty, \\ \sup_{N \geq 2} H_N(F^N(0)) &< \infty. \end{aligned}$$

Then, both the particle system (26) and the 2D Navier–Stokes Eq. (4) are globally well-posed and $(F^N(t))_{N \geq 2}$ is f_t -chaotic.

If we assume furthermore that initially, $(F^N(0))$ is f_0 -entropy chaotic as in Definition 5, then $(F^N(t))_{N \geq 2}$ is also f_t -entropy chaotic and $F_k^N \rightarrow f_t^{\otimes k}$ strongly in $L^1((\mathbb{R}^2)^k)$.

The proof of Theorem 7 follows the classical tightness/consistency/uniqueness arguments. The dissipation of the entropy

$$H_N(F^N) + \sigma \int_0^t I_N(F^N(s)) ds = H_N(F^N(0)),$$

and a control on a moment of $F^N(t)$ will give us a bound on

$$\int_0^T I_N(F^N(t)) dt.$$

This will in turn essentially bound several quantities for instance

$$\mathbb{E} \left[\sup_{0 < s < t < T} \frac{|X_i(t) - X_i(s)|}{|t - s|^\alpha} \right],$$

which finally helps to complete the tightness/consistency argument. The uniqueness argument is based on [2].

3.4 Other Extensions

We only very briefly mention some other recent extensions to the classical theory.

The first such case concerns kernels which are only locally Lipschitz. Such models have been studied in the context of flocking models, in [7] for example, and models of neuron dynamics, see [11] for instance, where the model is even more general with the diffusion coefficients possibly depending on the law.

Those models include interaction terms that are all locally Lipschitz but with a Lipschitz constant which grows to infinity when the region considered grows to the whole \mathbb{R}^d .

The classical method has to be adapted with typically a faster decay assumed on the limit law f_t . One key ingredient in the proof is then to show that trajectories cannot escape to infinity, typically because the model includes confining forces. In the absence of such assumptions, the problem can become ill-posed as shown in [70].

Finally, we mention that in the coming article [44], a new coupling strategy and a Glivenko–Cantelli theorem are used to show the mean field limits for systems (1) or (3) with global Hölder continuous interaction kernels $K \in C^{0,\alpha}$. For first-order system, $\alpha > 0$ is enough. But it requires $\alpha > \frac{1}{3}$ for second-order systems in order to ensure the existence of a differentiable stochastic flow.

4 A New Statistical Approach

In the authors' recent articles [47] and [48], we proposed a new relative entropy method to deal with mean field limit for very rough interaction kernels K .

The idea is to directly compare the distance between the joint distribution $F^N(t)$ solving the Liouville equation (12) and the tensor product of the limit law $f_t^{\otimes N}$ through the relative entropy

$$H_N(F^N | f^{\otimes N}) = \frac{1}{N} \int_{E^N} F^N \log \frac{F^N}{f^{\otimes N}} dz_1 \cdots dz_N.$$

The main theorem for the second-order systems with $\sigma > 0$ in [47] can be stated simply as follows.

Theorem 8 (Main Theorem in [47]) *Assume that $K \in L^\infty$ and that there exists $f_t \in L^\infty([0, T], L^1(E) \cap W^{1,p}(E))$ for every $1 \leq p \leq \infty$ which solves the limiting equation (2) with in addition*

$$\theta_f = \sup_{t \in [0, T]} \int_{\Gamma \times \mathbb{R}^d} e^{\lambda_f |\nabla_v \log f_t|} f_t dx dv < \infty,$$

for some $\lambda_f > 0$. Furthermore, assume initially that

$$\sup_{N \geq 2} H_N(F^N(0)) < \infty, \quad H_N(F^N(0) | f_0^{\otimes N}) \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

and

$$\sup_{N \geq 2} \frac{1}{N} \int_{E^N} \sum_{i=1}^N (1 + |z_i|^2) F^N(0, z_1, \dots, z_N) dz_1 \cdots dz_N < \infty.$$

Then, there exists a universal constant $C > 0$ s.t. for any $t \in [0, T]$,

$$H_N(F^N(t)|f_t^{\otimes N}) \leq e^{C \|K\|_{L^\infty} \theta_f t / \lambda_f} \left(H_N(F^N(0)|f_0^{\otimes N}) + \frac{C}{N} \right).$$

This result implies a strong form of propagation of chaos as the k -marginal converges to $f^{\otimes k}$ in L^1 . Indeed, if for instance $H_N(F^N(0)|f_0^{\otimes N}) \lesssim N^{-1}$, then the classical Csiszár–Kullback–Pinsker inequality (see Chapter 22 in [73] for instance) implies that

$$\|F_k^N(t) - f_t^{\otimes k}\|_{L^1} \leq \sqrt{2k H_k(F_k^N(t)|f_t^{\otimes k})} \lesssim \frac{1}{\sqrt{N}}.$$

The argument is in essence a weak–strong estimate comparing a very smooth solution to the limiting equation with a weak solution F_t^N to the Liouville equation (12). The heart of the proof consists of precise combinatoric estimates which lead to a new type of law of large numbers.

In a coming article [48], we extend the result to the first-order system (3) with $K \in W^{-1,\infty}$, i.e., K is the derivative of a bounded function but with the restriction that $\operatorname{div}_x K \in L^\infty$. By a careful truncation of the Biot–Savart kernel K and repeating our procedure, we can also provide an explicit convergence rate for stochastic vortex model (26) approximating 2D Navier–Stokes equation.

References

1. Ahn, S. M., Ha, S.-Y.: Stochastic flocking dynamics of the Cucker-Smale model with multiplicative white noise. *J. Math. Physics*, **51**, 103301 (2010)
2. Ben-Artzi, M.: Global solutions of two-dimensional Navier-Stokes and Euler equations. *Arch. Rational Mech. Anal.* **128**, 329–358 (1994)
3. Benedetto, D., Caglioti, E., Carrillo, J.A., Pulvirenti, M.: A non Maxwellian steady distribution for one-dimensional granular media. *J. Stat. Phys.* **91**, 979–990 (1998)
4. Bodineau, T., Gallagher, I., Saint-Raymond, L.: The Brownian motion as the limit of a deterministic system of hard-spheres. *Invent. Math.* **203**, 493–553 (2016)
5. T. Bodineau, T., Gallagher, I., Saint-Raymond, L.: From hard spheres dynamics to the Stokes-Fourier equations: an L^2 analysis of the Boltzmann-Grad limit. *C. R. Math. Acad. Sci. Paris* **353**, 623–627 (2015)
6. Bogoliubov, N. N.: Kinetic equations. *Journal of Physics USSR* **10**, 265–274 (1946)
7. Bolley, F., Cañizo, J. A., Carrillo, J. A.: Stochastic mean-field limit: non-Lipschitz forces and swarming. *Math. Mod. Meth. App. S.* **21**, 2179–2210 (2011)
8. Bolley, F., Cañizo, J. A., Carrillo, J. A.: Mean-field limit for the stochastic Vicsek model. *Appl. Math. Lett.* **25**, 339–343 (2012)
9. Bolley, F., Guillin, A., Malrieu, F.: Trend to equilibrium and particle approximation for a weakly self-consistent Vlasov-Fokker-Planck equation. *Math. Model. Numer. Anal.* **44**, 867–884 (2010)

10. Bolley, F., Guillin, A., Villani, C.: Quantitative concentration inequalities for empirical measures on non-compact space. *Probab. Theory Relat. Fields* **137**, 541-593 (2007)
11. Bossy, M., Faugeras, O., Talay, D.: Clarification and complement to “Mean-field description and propagation of chaos in networks of Hodgkin-Huxley and FitzHugh-Nagumo neurons”. *J. Math. Neurosci.* **5** Art. 19, 23 pp. (2015)
12. Bossy, M., Jabir, J. F., Talay, D.: On conditional McKean Lagrangian stochastic models. *Probab. Theory Relat. Fields* **151**, 319-351 (2011)
13. Cattiaux, P., Guillin, A., Malrieu, F.: Probabilistic approach for granular media equations in the non-uniformly convex case. *Probab. Theory Relat. Fields* **140**, 19-40 (2008)
14. Caglioti, E., Lions, P.L., Marchioro, C., Pulvirenti, M.: A special class of two-dimensional Euler Equations: A statistical mechanics description. *Commun. Math. Phys.* **143**, 501-525 (1992)
15. Caglioti, E., Lions, P.L., Marchioro, C., Pulvirenti, M.: A special class of two-dimensional Euler Equations: A statistical mechanics description. Part II. *Commun. Math. Phys.* **174**, 229-260 (1995)
16. Carlen, E.A., Carvalho, M.C., Le Roux, J., Loss, M., Villani, C.: Entropy and chaos in the Kac model. *Kinet. Relat. Models* **3**, 85-122 (2010)
17. Carrillo, J. A., Choi, Y.-P., Hauray, M.: The derivation of swarming models: Mean Field limit and Wasserstein distances. In: *Collective Dynamics from Bacteria to Crowds*, volume 553 of CISM International Centre for Mechanical Sciences, pages 1-46. Springer Vienna, (2014)
18. Carrillo, J. A., Fornasier, M., Toscani, G., Vecil, F.: Particle, kinetic, and hydrodynamic models of swarming. In: *Mathematical modeling of collective behavior in socio-economic and life sciences*. pp. 297-336, Model. Simul. Sci. Eng. Technol., Birkhauser Boston, Inc., Boston, MA, (2010)
19. Carrillo, J. A., DiFrancesco, M., Figalli, A., Laurent, T., Slepcev, D.: Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.* **156**, 229-271 (2011)
20. Carrillo, J. A., Lisini, S., Mainini, E.: Gradient flows for non-smooth interaction potentials. *Nonlinear Anal.* **100** 122-147 (2014)
21. Chuang, Y.L., Huang, Y.R., D'Orsogna, M.R., Bertozzi, A.L.: Multi-vehicle flocking: scalability of cooperative control algorithms using pairwise potentials. *IEEE Int. Conf. Robotics. Automation*, 2292-2299 (2007)
22. Cépa, E., Lépinig, D.: Diffusing particles with electrostatic repulsion. *Probab. Theory Relat. Fields* **107**, 429-449 (1997)
23. Cercignani,C., Illner, R., Pulvirenti, M.: *The Mathematical Theory of Dilute Gases*. Springer-Verlag, New York, (1994)
24. Cucker, F., Smale, S.: Emergent behavior in flocks. *IEEE Trans. Automat. Control* **52**, 852-862 (2007)
25. Degond, P., Frouvelle, A., Liu, J.-G.: Macroscopic limits and phase transition in a system of self-propelled particles. *J. Nonlinear Sci.* **23** 427-456 (2013)
26. Desvillettes, L., Graham, C., Méléard, S.: Probabilistic interpretation and numerical approximation of a Kac equation without cutoff. *Stochastic Process. Appl.* **84**, 115-135 (1999)
27. Dirr, N., Stamatakis, M., Zimmer, J.: Entropic and gradient flow formulations for nonlinear diffusion. *ArXiv: 1508.00549* (2016)
28. Flandolia, F., Gubinelli, M., Priolac, E.: Full well-posedness of point vortex dynamics corresponding to stochastic 2D Euler equations. *Stoch. Process. Appl.* **121**, 1445-1463 (2011)
29. Fournier, N., Hauray, M., Mischler, S.: Propagation of chaos for the 2d viscous vortex model. *J. Eur. Math. Soc.* **16**, 1425-1466 (2014)
30. Fournier, N., Jourdain, B.: Stochastic particle approximation of the Keller-Segel Equation and two-dimensional generalization of Bessel process. *ArXiv: 1507.01087* (2015)
31. Gallagher, I., Saint-Raymond, L., Texier, B.: From newton to Boltzmann: hard spheres and short-range potentials. In: *Zurich Advanced Lectures in Mathematics Series*, (2014)
32. Gibbs, J. W.: On the Fundamental Formulae of Dynamics. *Amer. J. Math.* **2** 49-64 (1879)
33. Gibbs, J. W.: Elementary principles in statistical mechanics: developed with especial reference to the rational foundation of thermodynamics. Dover publications, Inc., New York, (1960)

34. Godinho, D., Quininao, C.: Propagation of chaos for a sub-critical Keller-Segel Model. *Ann. Inst. H. Poincaré Probab. Statist.* **51**, 965-992 (2015)
35. Golse, F.: On the dynamics of large particle systems in the mean field limit. In: *Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity*. Volume 3 of the series *Lecture Notes in Applied Mathematics and Mechanics*, pp. 1-144. Springer, (2016)
36. Grad, H.: On the kinetic theory of rarefied gases. *Comm. on Pure and Appl. Math.* **2**, 331-407 (1949)
37. Graham, C., Méléard, S.: Stochastic particle approximations for generalized Boltzmann models and convergence estimates. *Ann. Probab.* **25**, 115-132 (1997)
38. Graham, C., Méléard, S.: Existence and regularity of a solution of a Kac equation without cutoff using the stochastic calculus of variations. *Comm. Math. Phys.* **205**, 551-569 (1999)
39. Ha, S.-Y., Lee, K., Levy, D.: Emergence of time-asymptotic flocking in a stochastic Cucker-Smale system. *Commun. Math. Sci.* **7**, 453-469 (2009)
40. Ha, S.-Y., Tadmor, E.: From particle to kinetic and hydrodynamic description of flocking. *Kinet. Relat. Models* **1**, 415-435 (2008)
41. Hauray, M., Mischler, S.: On Kac's chaos and related problems. *J. Funct. Anal.* **266**, 6055-6157 (2014)
42. Hauray, M., Jabin, P. E.: Particle Approximation of Vlasov Equations with Singular Forces. *Ann. Scient. Ecole Norm. Sup.* **48** 891-940 (2015)
43. Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation (JASSS)*. **5**, no. 3, (2002)
44. Holding, T.: Propagation of chaos for Hölder continuous interaction kernels via Glivenko-Cantelli. Manuscript, Personal communication.
45. K. Itô: On stochastic differential equations. *Memoirs of the American Mathematical Society*. **4**, 1-51 (1951)
46. Jabin, P.E.: A review for the mean field limit for Vlasov equations. *Kinet. Relat. Models* **7**, 661-711 (2014)
47. Jabin, P.E., Wang, Z.: Mean field limit and propagation of chaos for Vlasov systems with bounded forces. [ArXiv: 1511.03769](https://arxiv.org/abs/1511.03769) (2015)
48. Jabin, P.E., Wang, Z.: Mean filed limit for stochastic 1st order systems with almost bounded stream functions. In preparation.
49. Jeans, J. H.: On the theory of star-streaming and the structure of the universe. *Monthly Notices of the Royal Astronomical Society* **76**, 70-84 (1915)
50. Kac, M.: Foundations of kinetic theory. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954-1955*, Vol. III, pp. 171-197. University of California Press, Berkeley (1956)
51. Kipnis, C., Landim, C.: Scaling limit of interacting particle systems. In: *Grundlehren der mathematischen Wissenschaften* 320. Springer, (1999)
52. Krause, U.: A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, pp. 227-236 (2000)
53. Lanford, O. E. III: Time evolution of large classical systems. In *Dynamical systems, theory and applications (Recontres, Battelle Res. Inst., Seattle, Wash., 1974)*, pp 1-111. Lecture Notes in Phys., Vol. 38. Springer, Berlin, (1975)
54. Lazarovici, D: The Vlasov-Poisson dynamics as the mean-field limit of rigid charges. [ArXiv:1502.07047](https://arxiv.org/abs/1502.07047) (2015)
55. Lazarovici, D., Pickl, P.: A Mean-field limit for the Vlasov-Poisson system. [ArXiv 1502.04608](https://arxiv.org/abs/1502.04608) (2015)
56. Liu, J.-G., Yang, R.: A random particle blob method for the Keller-Segel equation and convergence analysis. *Math. Comp.*, to appear.
57. Malrieu, F.: Logarithmic Sobolev inequalities for some nonlinear PDE's. *Stoch. Process. Appl.* **95**, 109-132 (2001)
58. Malrieu, F.: Convergence to equilibrium for granular media equations and their Euler schemes. *Ann. Appl. Probab.* **13**, 540-560 (2003)

59. Marchioro, C., Pulvirenti, M.: Hydrodynamics in two dimensions and vortex theory. *Commun. Math. Phys.* **84**, 483-503 (1982)
60. McKean, H.P. Jr.: Propagation of chaos for a class of non-linear parabolic equations. In: Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967), pp. 41-57. Air Force Office Sci. Res., Arlington, VA, (1967)
61. Méléard, S.: Asymptotic behavior of some interacting particle systems; McKean-Vlasov and Boltzmann models. In: Probabilistic Models for Nonlinear Partial Differential Equations (Lecture Notes in Mathematics), Vol. 1627, Springer, (1996)
62. Méléard, S.: Monte-Carlo approximation for 2d Navier-Stokes equations with measure initial data. *Probab. Theory Relat. Fields* **121**, 367-388 (2001)
63. Mischler, S., Mouhot, C.: Kac's Program in Kinetic Theory. *Invent. Math.* **193**, 1-147 (2013)
64. Mischler, S., Mouhot, C., Wennberg, B.: A new approach to quantitative chaos propagation for drift, diffusion and jump process. *Probab. Theory Relat. Fields* **161**, 1-59 (2015)
65. Motsch, S., Tadmor, E.: A new model for self-organized dynamics and its flocking behavior. *J. Stat. Phys.*, **144**, 923-947 (2011)
66. Osada, H.: A stochastic differential equation arising from the vortex problem. *Proc. Japan Acad. Ser. A Math. Sci.* **62**, 333-336 (1986)
67. Osada, H.: Propagation of chaos for the two-dimensional Navier-Stokes equation. In: Probabilistic methods in mathematical physics (Katata/Kyoto, 1985), pp. 303-334. Academic Press, Boston, MA, (1987)
68. Othmer, H. G., Stevens, A.: Aggregation, blowup, and collapse: the ABCs of taxis in reinforced random walks. *SIAM J. Appl. Math.* **57**, 1044-1081 (1997)
69. Perthame, B.: Transport equations in biology. Frontiers in Mathematics. Birkhäuser Verlag, Basel, (2007)
70. Scheutzow, M.: Uniqueness and non-uniqueness of solutions of Vlasov-McKean equations. *J. Austral. Math. Soc. Series A*, **43**, 246-256 (1987)
71. Sznitman, A.-S.: Topics in propagation of chaos. In: Ecole d'été de probabilités de Saint-Flour XIX-1989, pp. 165-251. Springer, (1991)
72. Topaz, C. M., Bertozzi, A. L., Lewis, M. A.: A nonlocal continuum model for biological aggregation. *Bull. Math. Biol.* **68**, 1601-1623 (2006)
73. Villani, C.: Optimal Transport, Old and New. In: Grundlehren der mathematischen Wissenschaften 338. Springer Science & Business Media, (2008)
74. Vicsek, T., Czirók, E., Ben-Jacob, E., Cohen, I., Shochet, O.: Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**, 1226-1229 (1995)
75. Vlasov, A. A.: On vibration properties of electron gas. *J. Exp. Theor. Phys.* (in Russian), **8** (3):291, (1938)
76. Vlasov, A. A.: The vibrational properties of an electron gas. *Sov. Phys. Usp.* **10**, 721-733 (1968)
77. Xia, H., Wang, H., Xuan, Z.: Opinion dynamics: A multidisciplinary review and perspective on future research. *International Journal of Knowledge and Systems Science (IJKSS)* **2**, 72-91 (2011)