Modeling and Simulation in Science,
Engineering and Technology

Nicola Bellomo
José Antonio Carrillo
Eitan Tadmor
Editors

Active Particles, Volume 3

Advances in Theory, Models, and Applications

Birkhäuser

Birkhäuser

Modeling and Simulation in Science, Engineering and Technology

More information about this series at https://link.springer.com/bookseries/4960

Nicola Bellomo • José Antonio Carrillo
Eitan Tadmor
Editors

Active Particles, Volume 3

Advances in Theory, Models, and Applications

Birkhäuser

Editors
Nicola Bellomo
Department of Mathematical Sciences
Politecnico di Torino, Turin, Italy

University of Granada, Granada, Spain

IMATI CNR, Pavia, Italy

José Antonio Carrillo
Mathematical Institute
University of Oxford
Oxford, UK

Eitan Tadmor
Department of Mathematics and
Institute for Physical Science & Technology
University of Maryland
College Park, MD, USA

ISSN 2164-3679 ISSN 2164-3725 (electronic)
Modeling and Simulation in Science, Engineering and Technology
ISBN 978-3-030-93301-2 ISBN 978-3-030-93302-9 (eBook)
https://doi.org/10.1007/978-3-030-93302-9

Mathematics Subject Classification: 74A25, 70F45, 82D99, 82B21, 35Q92, 35Q70, 37N40, 49J15

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This edited book collects six surveys on modeling, qualitative analysis, and simulation of active matter focusing on specific applications in natural sciences. It is a follow-up to volumes 1 and 2 under the same title. The book, as in the preceding volumes, blends together contributions which indicate the diversity of the subject matter in theory and applications within an interdisciplinary framework which requires the use of different mathematical tools. Indeed, this new frontier of science offers a range of new challenging problems which requires advanced mathematical tools and, in some cases, new mathematical theories. The contents are as follows:

Chapter “Variability and Heterogeneity in Natural Swarms: Experiments and Modeling,” by Ariel, Ayali, Be’er, and Knebel, focuses on some of the fundamental properties of heterogeneous collectives in nature, with an emphasis on two widely used model organisms—swarming bacteria and locusts. The aim consists in explaining the observed phenomena in view of laboratory experiments. Surprisingly, the authors observe that while heterogeneity typically discourages collectivity, there are several natural examples where it has the opposite effect. A detailed study of heterogeneity is a key feature of this chapter.

Chapter “Active Crowds,” by Bruna, Burger, Pietschmann, and Wolfram, enlightens the multiscale aspects of the dynamics of human crowds, where heterogeneity is considered as a common feature of all living systems. The chapter is devoted to show how macroscopic models can be derived from the underlying description at the microscopic scale by tools inspired by methods of statistical physics.

Chapter “Mathematical Modeling of Cell Collective Motion Triggered by Self-Generated Gradients,” by Calvez, Demircigil, and Sublet, develops a robust strategy to model how a group of cells find its way during a long journey. Various scenarios for modeling traveling waves are studied for cells that constantly deplete a chemical cue, and so create their own signaling gradient all along the way are studied. Analytic problems refer also to the celebrated model by Keller and Segel for bacterial chemotaxis.

Chapter “Clustering Dynamics on Graphs: From Spectral Clustering to Mean Shift Through Fokker-Planck Interpolation,” by Craig, García Trillos, and Slepčev, proposes a unifying framework to interpolate between density-driven and geometry-based algorithms for data clustering and, specifically, to connect the mean shift algorithm with spectral clustering at discrete and continuum levels. New forms of mean shift algorithms on graphs provide a new theoretical insight on the behavior of the family of diffusion maps in the large sample limit.

Chapter "Random Batch Methods for Classical and Quantum Interacting Particle Systems and Statistical Samplings," by Jin and Li, deals with the random batch methods for interacting particle systems with the aim of reducing the computational cost. These methods are referred to both classical and quantum systems, the corresponding theory, and applications from molecular dynamics and statistical samplings to agent-based models for collective behavior and quantum Monte-Carlo methods.

Chapter "Trends in Consensus-Based Optimization," by Totzeck, delivers an overview of the consensus-based global optimization algorithm and its recent variants. The contents start from the formulation and analytical results of the original model, then the focus moves to variants using component-wise independent or common noise. The authors discuss the relationship of consensus-based optimization with particle swarm optimization, a method widely used in the engineering community.

Finally, we mention the project Ki-Net—an NSF Research Network focused on "Kinetic description of emerging challenges in multiscale problems of natural sciences" (www.ki-net.umd.edu)— fostered a series of activities with main intellectual focus on development, analysis, computation, and application of quantum dynamics, network dynamics, and kinetic models of biological processes. Therefore, the project contributed to the research activity during the editing of volumes 1 and 2. The scientific legacy of the Ki-Net project is still a strong motivation to develop the research activity in the field as witnessed by the chapters which have contributed to this book.

Turin, Italy	Nicola Bellomo
Oxford, UK	José Antonio Carrillo
College Park, MD, USA	Eitan Tadmor

Contents

Variability and Heterogeneity in Natural Swarms: Experiments and Modeling

G. Ariel, A. Ayali, A. Be'er, and D. Knebel

Abstract Collective motion of large-scale natural swarms, such as moving animal groups or expanding bacterial colonies, has been described as self-organized phenomena. Thus, it is clear that the observed macroscopic, coarse-grained swarm dynamics depend on the properties of the individuals of which it is composed. In nature, individuals are never identical and may differ in practically every parameter. Hence, intragroup variability and its effect on the ability to form coordinated motion is of interest, both from theoretical and biological points of view. This review examines some of the fundamental properties of heterogeneous collectives in nature, with an emphasis on two widely used model organisms: swarming bacteria and locusts. Theoretical attempts to explain the observed phenomena are discussed in view of laboratory experiments, highlighting their successes and failures. In particular we show that, surprisingly, while heterogeneity typically discourages collectivity, there are several natural examples where it has the opposite effect.

G. Ariel (✉)
Department of Mathematics, Bar-Ilan University, Ramat-Gan, Israel
e-mail: arielg@math.biu.ac.il

A. Ayali
School of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
e-mail: ayali@tauex.tau.ac.il

A. Be'er
Zuckerberg Institute for Water Research, The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Midreshet Ben-Gurion, Israel

Department of Physics, Ben-Gurion University of the Negev, Beer-Sheva, Israel

D. Knebel
School of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel

Lise Meitner Group Social Behaviour, Max Planck Institute for Chemical Ecology, Jena, Germany

N. Bellomo et al. (eds.), *Active Particles, Volume 3*, Modeling and Simulation in
Science, Engineering and Technology, https://doi.org/10.1007/978-3-030-93302-9_1

1 Introduction

Collective behavior is ubiquitous in living organisms at all levels of complexity. An important type of collective behavior is the translocation of groups, known as collective motion (Krause et al. 2002). Here, we refer to collective motion as macroscopic, synchronized, or coordinated movement of individuals that arises from small-scale, local inter-individual interactions (Giardina 2008; Sumpter 2010; Vicsek and Zafeiris 2012). Collective motion is found in the context of foraging for food, shelter seeking, or predator evasion, but also in other, less clearly defined or recognized circumstances (Herbert-Read et al. 2017). As noted, it can be found in practically all phylogenetic groups, from single cells (Be'er and Ariel 2019; Schumacher et al. 2016) to humans (Barnett et al. 2016; Castellano et al. 2009; Faria et al. 2010; Helbing 2001), as well as in synthetic entities like simulated agents (Kennedy and Eberhart 1995), self-propelled inanimate particles (Bär et al. 2020), or motile robots (Dorigo et al. 2020).

More than three decades ago, the phenomenon of collective motion has been described as *emergent* and *self-organizing* (Ben-Jacob et al. 2000; Vicsek et al. 1995; Vicsek and Zafeiris 2012), i.e., the congruence of local interaction to macroscopic, group-level dynamics. The field has evolved into an active, interdisciplinary research field, comprising physicists and mathematicians, computer scientists, engineers, and biologists; all trying to identify principles that are fundamental to the self-organized emergent phenomenon and its intricate connection to movement and migration. The key questions that are common to collective motion research are related to the identification of interactions between the individual, the collective, and the environment and to understanding how these converge into coherent synchronized motion (e.g., Ariel and Ayali 2015; Bär et al. 2020; Couzin et al. 2005; Edelstein-Keshet 2001; Giardina 2008; Tadmor 2021; Vicsek and Zafeiris 2012).

These questions have attracted renewed interest in light of recent technological advances, in particular computer-vision based tracking methods. This has spurred intense experimental research of collective motion under controlled laboratory environment, which lent itself to quantitative analysis of individual and crowd movement. For example, automated individual tracking systems based on body-marks recognition or on miniature barcodes allow continuous and consistent simultaneous high-precision monitoring of all individuals in animal groups, in an attempt to decipher the intricate underlying interactions. In addition, new methods allow collecting data on the movements and interactions of multiple animals in their natural environmental setting. This facilitates testing interactions of the collective with complex environments.

Much of the progress that has been made in our understanding of the causes and consequences of collective motion has been gained by comparing experimental observations with mathematical and computational models. At the same time, ample theoretical work on collective motion includes a wide range of theoretical approaches, suggesting explanations for the emergence of collective motion, its robustness and evolutionary advantages (e.g., Ariel and Ayali 2015; Be'er and Ariel

2019; Carrillo et al. 2010; Degond and Motsch 2008; Giardina 2008; Ha and Tadmor 2008; Tadmor 2021; Toner et al. 2005; Wensink et al. 2012 and the references therein).

In general, modeling approaches can be categorized as either continuous models, written in terms of integro-differential equations, or discrete agent-based models (ABMs). Continuous models typically describe the coarse-grained density of animals and other system constituents as continuous fields, e.g., by coupled reaction-diffusion equations or, following a kinetic approach, by hydrodynamic or Boltzmann equations (Ben-Jacob et al. 2000; Carrillo et al. 2010; Degond and Motsch 2008; Edelstein-Keshet 2001; Ha and Tadmor 2008; Tadmor 2021; Toner et al. 2005; Wensink et al. 2012). One of the main drawbacks of continuous models is the difficultly of relating actual properties of individual animals (e.g., body shape and size, hunger, and other internal states) to specific details of the model (Edelstein-Keshet 2001). This is one of the reasons why much of the current theoretical work related to collective motion of *real* animal experiments comprise agent-based simulations, which are useful for generating the dynamics from the point of view of the individual animal ("Umwelt" in biology or "Lagrangian description" in physics). The dynamics in ABMs, also referred to as self-propelled particles (SPPs), are given by specifying the internal state of each animal, its interaction with others (conspecifics), and its interactions with the environment (Bär et al. 2020; Edelstein-Keshet 2001; Giardina 2008). However, such models are limited by the number of agents that can be simulated due to computational capacities (very far, for example, from the millions of individuals comprising a locust swarm or trillions of cells in a bacterial colony). In addition, they do not provide a macroscopic, or coarse-grained, description of the swarm dynamics as a whole, and additional mathematical tools may be needed to interpret the results. Accordingly, the theoretical modeling of coarse-graining ABMs is a highly challenging research topic (Carrillo et al. 2010; Degond and Motsch 2008; Ha and Tadmor 2008; Ihle 2011; Toner et al. 2005; Wensink et al. 2012).

One important aspect of collective motion research, experimental and theoretical alike, is that of the level of similarity between the individuals composing the group, also referred to as the group homogeneity. Collective motion requires consensus in the sense that individuals need to adjust their behavior according to conspecifics. In other words, it is expected that collectivity will result in some homogenization among the individuals forming the group. At the same time, the group dynamics should somehow be a function of its constituents, i.e., depend on the different traits of the individuals composing it. A group can be heterogeneous at many different levels, including permanent differences or transient ones (e.g., due to different interactions with the surroundings), as will be discussed in the following sections. This heterogeneity can have important consequences on collective motion, leading to distinct group properties and variability between different groups (del Mar Delgado et al. 2018; Herbert-Read et al. 2013; Jolles et al. 2020; Knebel et al. 2019; May 1974; Ward et al. 2018).

These general statements bring about ample open questions that are related to the cross-dependency between individual heterogeneity and collective motion (Giardina

2008; Sumpter 2010; Vicsek and Zafeiris 2012). For example, which traits of the individual are adjusted in order for it to become part of the synchronized group? On the other direction, it is not clear whether variability supports or interferes with collectivity, and under what circumstances is heterogeneity biologically favorable? One of the main goals of this review is to develop a general methodology for addressing these issues and its application to experiments. Deciphering the bi-directional interactions between individual and group properties is essential for understanding the swarm phenomenon and predicting large-scale swarm behaviors.

To this end, we begin with a review of the different *sources* of variability in biology, relevant to collective motion (Sect. 2). Section 3 describes the *impact* of variability on collective motion as observed in experiments. We will see that, while variability is typically a limiting factor for collectivity, in some cases it may enhance it. Moreover, reduced order is not always a disadvantage. Section 4 surveys the literature on modeling heterogeneous collective motion, providing a historical overview, spanning some 50 years of progress. A few examples comparing theoretical predictions with experiments will be discussed. We conclude in Sect. 5 with our own perspective on interesting directions for future research.

2 Sources of Variability in Nature

Variability is a key concept in biology. Whether structural, functional, or behavioral, variability among animals and within an individual along time is essential for adaptability to the environment and for survival. One important aspect in which variability plays a dominant role is in the context of collective behavior, in particular during movement. Both permanent and transient differences among and within animals may be instrumental in the dynamics and organization of the group, ranging from local interactions between conspecifics to macroscopic organization. In this section, we outline several biological sources of variability that affect collective motion.

2.1 Development as a Source of Variation

All animals change and develop during their lifetime. Ultimately, through this process, individuals go from immature early stages to being able to reproduce and give rise to surviving offspring. Yet, the degree of such changes differs greatly among taxa. As the time scale of developmental changes is typically longer than the characteristic time scale in which swarming occurs or is observed, groups that are composed of individuals at different developmental levels are intrinsically heterogeneous.

For example, insects can be classified into two major groups according to their ontogeny: Holometabola, in which the insect goes through an extreme metamorphosis during its development; and hemimetabola, in which the changes between immature and mature individuals are milder.

Holometabola insects go through several distinct life stages, differing in their anatomy and morphology, as well as in physiology and behavior: the larva (hatching from eggs), pupa and imago (adult insect). Among the Holometabola one finds many of the truly social (called eusocial) insects, i.e., insects that live in cooperative colonies such as bees and ants. Their collective behavior is complex, both on the inter-individual communication level and the exhibited behaviors.

Hemimetabola insects go through a series of larval stages. The basic anatomy and many features of the behavior of the larvae (or nymphs) and adults are rather similar (except for flight and reproduction related ones). Locusts are one of the prominent examples of hemimetabola insects exhibiting collective motion, notorious for forming swarms composed of millions of individuals. The non-adult (non-flying) insects migrate in huge marching bands. These often include nymphs of different developmental stages or different larval instars, thus introducing many aspects of variation to the group, for example, in body size, walking speed, and food consumption (Ariel and Ayali 2015). To the best of our knowledge, no research specifically addressed the influence of instar variance (i.e., swarms of nymphs at a mixture of developmental stages) upon the swarm's dynamics.

Fish go through continuous development. They hatch from the egg into a larva state, characterized by the ability for exogenous (external) feeding. Next, fish go through juvenile phases, in which the body structure changes, and eventually mature into sexually active adult fish (the exact definition of these stages is ambiguous; see Penaz 2001). During this development, fish grow and change their behavior. The developmental level is critical to the formation of schools, as part of the behavioral change, which is highly relevant to the collective motion of the school or shoal. In particular, the level of attraction to conspecifics was shown to increase during juvenile development (Hinz and de Polavieja 2017).

Mammals show a very distinctive maturation that includes no metamorphosis. Offspring are born in rather small batches and are highly dependent on their mother for feeding and protection. As a result, collective behavior in mammals includes co-behavior of several generations at once. This inter-generational composition might be instrumental for understanding animal packs (Ákos et al. 2014; Leca et al. 2003; McComb et al. 2011; Strandburg-Peshkin et al. 2015, 2017) and human crowds (Barnett et al. 2016; Faria et al. 2010).

2.2 *Transient Changes in the Behavior of Individuals*

Rarely, if at all, will a moving animal maintain constant dynamics on the go. Such changes include speed, switching between moving and pausing, and more. When moving in a group, individual kinematic changes increase the propensity for variability within the group, and thus, essentially add noise to a synchronized collective. However, such temporal variations at the individual level may also contribute to the overall movement and success of group-level tasks (Viscido et al. 2004).

For example, locusts walk in a pause-and-go motion pattern (Ariel et al. 2014; Bazazi et al. 2012), i.e., they intermittently switch between walking and standing. The durations of walking and pausing bouts have different distributions: while walking bout durations are approximately exponentially distributed, pauses show an approximate power-law tail. This indicates that while the termination of walking bouts reflects a random, memory-less process (Reches et al. 2019), the termination of pauses is based on information processing with a memory (Ariel et al. 2014). What kind of information is being processed? Looking into the pause episodes of marching locusts, the termination of pauses was found to depend on the locusts' social environment. Both tactile stimulation and visual inputs make a locust stop standing and engage in walking. Thus, when a locust is touched by another locust or alternatively is seeing locusts depart from its front visual field or appear at its rear, its probability to start walking increases. Moreover, as locusts rarely turn during a walking bout, the shift from standing to walking is crucial for directional changes in order to align with the crowd (Ariel et al. 2014; Knebel et al. 2021).

2.3 Environmentally Induced Variations

The behavior of a moving organism is affected by external conditions, including the physical habitat (Strandburg-Peshkin et al. 2017), the ecological niche (Ward et al. 2018) and the topology of the environment (Amichay et al. 2016; May 1974; Strandburg-Peshkin et al. 2015). Therefore, differences in environmental characteristics can also exert different constraints on collective motion, inducing inter-environment variability. For example, predation is an environmental factor that can shape the behavioral strategies of many organisms. The abatement of predation risk, e.g., through predator confusion and increased vigilance, was suggested as one of the dominant advantages of aggregation (Ioannou et al. 2012; Krause et al. 2002). Thus, this may be a major evolutionary pressure leading to collective animal motion. The relation between predation and collective motion was demonstrated in fish. For example, fish that are grown in high predation-risk environments showed higher group cohesiveness (Herbert-Read et al. 2017). This exemplifies how the ecological niche can dictate group dynamics by modulating local individual decisions.

2.4 Social Structure

The social environment is another factor that affects variance (Smith et al. 2016; Ward and Webster 2016). The level of disparity in social rank among individuals can vary depending on the society being a complete egalitarian one, or one based on an hierarchical structure (Ákos et al. 2014; Couzin et al. 2005; Garland et al. 2018; Jacoby et al. 2016; Leca et al. 2003; Lewis et al. 2011; Nagy et al. 2010; Smith et al. 2016; Strandburg-Peshkin et al. 2015; Watts et al. 2017). Below we discuss a few natural examples.

In clonal raider ants, a queen-less species in which each ant can reproduce by parthenogenesis (reproduction without fertilization), the size of the colony dictates the division of labor structure (an organizational regime in which individual ants are assigned to different tasks in the colony). If the colony is small, then each ant fulfills various tasks, both within and outside the nest. However, in large colonies, different ants occupy different roles (Ulrich et al. 2018).

Flocks of pigeons exhibit outstanding flight synchronization. The flock is organized through a hierarchical network, where different individuals have different influence on other pigeons. It is estimated that such hierarchical structure, rather than egalitarian alternatives, is more efficient for coordinated flight in small flocks (Nagy et al. 2010).

Many birds use thermals to climb up, thereby reducing their costly need to flap their wings. Yet, the use of thermals can differ between individuals, as was shown for white storks that form flocks with leader–follower relations (Flack et al. 2018). Leaders tend to explore for thermals, while the followers enjoy their findings. Yet, followers exit the thermals earlier than leaders, rise less, and must flap their wings more. Thus, leader–follower social relations combine with environmental factors. On the one hand, shared knowledge decreases variation, while on the other, different exploitation of resources increases it.

Social structure may go beyond a linear ranking scale. Indeed, the social network of relatedness and familiarities can influence the stability of swarm dynamics and the organization within it (Barber and Ruxton 2000; Barber and Wright 2001; Croft et al. 2008).

2.5 *Inherent/Intrinsic Properties and Animal Personality*

Inherent variability among individuals is, perhaps, what makes biological systems essentially different from ideal theoretical models. Each biological "agent" is unique and has its own properties. These can be anatomical features like body size or physiological parameters such as metabolic rate. Individuals may also have personal behavioral characteristics that are consistent across different contexts, also referred to as animal personality (Wolf and Weissing 2012). For example, properties such as boldness, aggressiveness, activity level, and sociability were considered as behavioral tendencies that make up an animal personality in a range of organisms, from insects to mammals (Gosling 2001). Such features induce differences in the behavior and decisions of individuals, which are influential in the formation of collective behavior.

For example, feral guppies show consistent variance in their boldness, activity level, and sociability (Brown and Irving 2014). However, the exploratory behavior of the groups they form was found to be independent of the average personality characteristics of its members. Nonetheless, low exploratory behavior did correlate with the activity score of the least active member in the group. Conversely, high exploratory behavior correlated with the sociality rank of the most social member (Munson et al. 2021). Therefore, extreme personalities of single individuals are critical for the entire group.

Body size is a parameter that covers many anatomical and physiological measurements, such as body mass, volume, and muscularity. These, of course, affect movement kinematics but also the relative impact of individuals on others. For example, the order within groups of schooling fish has been shown to reflect the heterogeneity of the member's body size. Larger fish tend to occupy the front and edges of a school, while smaller ones populate the center and the back. Consequently, larger fish tend to have a higher influence on the group direction of movement (Jolles et al. 2020). Other experiments found different spatial distribution of body sizes within the swarm, depending on species (Romey 1997; Theodorakis 1989), suggesting that the effect of body size is coupled to other properties (Sih 1980).

2.6 Variability in Microorganisms

Microorganisms grow in nature in a variety of habitats, from aquatic niches and soil, to waste and within hosts. In much of these systems, several species, or variants of the same species, occupy the same niche, creating a heterogeneous population with a diverse range of interactions between them (Ben-Jacob et al. 2016). For example, *Bacillus subtilis* is a model organism used in swarm assays. Typical swarms of *B. subtilis* form a multilayered colonial structure composed of billions of cells. Grown from a single cell, colonies become a mixed population of two strikingly different cell types. In one type the transcription factor for motility is active, and in the other one motility is off and the bacteria are placed in long chains of immobile cells (Kearns and Losick 2005). Cell population heterogeneity could enable *B. subtilis* to exploit its present location through the production of immobile cells as well as to explore new environmental niches through the generation of cells with different motility capabilities, resistance to harmful substances, and response to chemical cues (Kearns and Losick 2005).

Multispecies communities cooperate and at the same time compete in order to survive harsh conditions (Ben-Jacob et al. 2016). Examples of experimental studies include biofilms (Nadell et al. 2016; Rosenberg et al. 2016; Tong et al. 2007), plant roots (Stefanic et al. 2015), neighboring colonies of *Bacillus subtilis* forming boundaries between non-kin colonies, swarming assays (showing either mixing or population segregation depending on species) (Tipping and Gibbs 2019), mixtures of motile and non-motile, antibiotic resistant species (Benisty et al. 2015; Ingham et al. 2011), and other collectively moving bacteria (Zuo and Wu 2020). In other works, it was shown that species diversity can lead to a non-transitive symbiosis in a "rock-paper-scissors" manner that leads to stable coexistence of all the species (Kerr et al. 2002; Reichenbach et al. 2007). Exploitation competition can lead to growth inhibition when one bacterial species changes its metabolic functions (Hibbing et al. 2010).

3 Experiments with Heterogeneous Swarms

Extensive experimental research has been devoted to understanding the effect of variability among individuals on the group's collective behavior—ranging from bacteria to primates (see Ben-Jacob et al. 2016; del Mar Delgado et al. 2018; Gosling 2001; Herbert-Read et al. 2013; Jolles et al. 2020; Ward and Webster 2016; Wolf and Weissing 2012 for recent reviews, and Dorigo et al. 2020 for investigation of heterogeneity in the context of swarm robotics). Indeed, it has been suggested that the inherent differences among members of the group can translate into distinct group characteristics (Brown and Irving 2014; Jolles et al. 2018; Knebel et al. 2019; Munson et al. 2021; Strandburg-Peshkin et al. 2017). Namely, different groups composed of individuals with distinctive features may adopt different collective behaviors. However, the interactions between variability in specific aspects of the individuals' behavior and group-level processes are complex and bi-directional. This leads to a practical difficulty in distinguishing between the inherent variability between individual features and the results of their interaction with the crowd. We stress that we focus on the interplay between variability and collective motion, not on other forms of collectivity, for example, shared resources or decision making.

Surprisingly, despite extensive research on the effect of heterogeneity on collective motion, general conclusions are scarce and simplistic. In some cases, the effect of heterogeneity is subtle and does not determine the movement of the group (Brown and Irving 2014). However, three main effects are generally accepted: First, collectivity reduces the inherent variability between individuals (Knebel et al. 2019; Planas-Sitjà et al. 2021). This is not surprising, as individuals are exposed to similar "averaged-out" environments. Second, heterogeneity quantitatively reduces order and synchronization (Jolles et al. 2017; Kotrschal et al. 2020; Ling et al. 2019). Note that reduced order is not necessarily disadvantageous. For example, it can assist in collective maneuvering around obstacles (Feinerman et al. 2018; Fonio et al. 2016; Gelblum et al. 2015) and enhance accurate sensing of the environment (Berdahl et al. 2013). Last, individual differences can determine the spatial organization within the swarm (Jolles et al. 2017). For example, faster individuals are typically at the front (Pettit et al. 2015).

Below, we focus on several examples, including a couple of exceptions going beyond these general conclusions.

3.1 Fish

Golden shiners are well known for their schooling behavior. In order to maintain group formation, individuals tend to stay at a small distance away from their closest neighbor, yet avoid proximity (approximately 1 body length (Katz et al. 2011)). They do so by adjusting their velocity according to the relative position and velocity of conspecifics. It has been found that fish swimming at higher speeds affect their neighbors to a greater extent. Thus, individual variation in speeds is instrumental in inter-fish interactions, serving as a key element in fish schooling (Jolles et al. 2017).

Guppy is a fish species with rather low shoaling behavior, in which different fish exhibit different collectivity tendencies. These tendencies, to some extent, pass from mothers to offspring. In a recent study (Kotrschal et al. 2020), selecting and breeding females who were swimming in high coordination with conspecifics led to increasingly higher collectivity scores. Therefore, individual tendencies to make appropriate social decisions might take role in the natural selection of swarming communities. Guppies also differ in individual shy-bold responses. It was shown that the composition of the small groups according to this trait affects foraging success (Dyer et al. 2009). In particular, groups that have bold fish find food faster. However, an all-bold school is not optimal as more fed fish are found in mixed groups. This observation was explained by the tendency of shy fish to follow bold ones and thus reach the food source immediately after them. Therefore, inter-individual heterogeneity can maximize groups' ability to use resources.

Experiments with giant danio showed that temporal variability in the speed and polarity leads to the emergence of several preferred collective states (Viscido et al. 2004).

3.2 Mammals

The effect of social structure on collective movement of mammals has been explored in several species (Smith et al. 2016), including monkeys (Leca et al. 2003; Strandburg-Peshkin et al. 2015, 2017), dolphins (Lewis et al. 2011), and family dogs (Ákos et al. 2014).

For example, baboons live in groups of up to 100 individuals, exhibiting multiple forms of collective behaviors, including collective motion (Strandburg-Peshkin et al. 2015, 2017). Unlike other, smaller animals, Baboons are studied in their natural habitat, imposing constraints on the ability to perform highly controlled experiments. Recent experiments applied high resolution GPS tracking (Strandburg-Peshkin et al. 2015) and unmanned aerial vehicle photography (Strandburg-Peshkin et al. 2017) to reconstruct animal trajectories in the wild. Some degrees of variations within the packs were found. For example, when on the move, different individuals show different preferred positions (central or peripheral) within the pack (Farine et al. 2017). The structure of the group and its navigational decisions were also shown to be highly dependent on the physical characteristics of the habitat and thereby change depending on the environment.

3.3 Insects

When an ant comes across goods which are too heavy for it to carry alone, it recruits other ants for assistance. Once a team is gathered, its members engage in a complex process of cooperative transport of the good to the nest. During this process, ants either pull or lift the item, but do not push. Thus, they arrange around it, with individuals facing the direction of movement lifting, while the ones on the opposite side pulling (Gelblum et al. 2015). As turns and angular modifications take place, ants may change their relative position in respect to the movement direction and switch roles, resulting in transient roles. These variations are essential for steering maneuvers and successful navigation to the nest. As carrying ants have limited sensing ability of the environment, they are assisted by freely moving ants around them. The latter, which are more knowledgeable about the path back to the nest, intermittently attach themselves to the load and pull in the required direction. However, their influence on the group is limited for a few seconds, after which other freely moving ants join the steering (Feinerman et al. 2018; Gelblum et al. 2015). Overall, individuals with different realizations of the environment participate in the collective effort, introducing small, cumulative changes to the direction of motion.

A plague of locusts can involve millions of individuals. Yet, under lab conditions, even a small group of 10 locusts can exhibit collective motion when placed in a ring-shaped arena. In such experiments, the group shows a consistent tendency to walk in either clockwise or counterclockwise direction with considerable agreement among the individuals (Buhl et al. 2006; Knebel et al. 2019). Despite this general formation of marching, different groups show different kinematic properties (e.g., the fraction of time spent walking and speed). Interestingly, while the differences among groups are significantly high, within each group (i.e., among the individuals) the differences are low. This indicates that each group develops a distinctive internal dynamic with specific kinematic features that are, on one hand, unique to the group, while on the other side practiced by all group members similarly (see Fig. 1). In (Knebel et al. 2019), it was shown that the origin of both the intergroup heterogeneity and the intragroup homogeny is in the individual socio-behavioral tendencies: different animals have different propensity of joining a crowd of walking conspecifics. Thus, the specific composition of locusts grouped together determines the specific dynamic the group eventually develops.

Not only are locust swarms intrinsically heterogeneous, recent laboratory experiments found that the connection between the properties of individuals changes fundamentally during collective motion. In Knebel et al. (2021), the walking kinematics of individual insects were monitored before, during, and after collective motion under controlled laboratory settings. It was found that taking part in collective motion induced unique behavioral kinematics compared with those

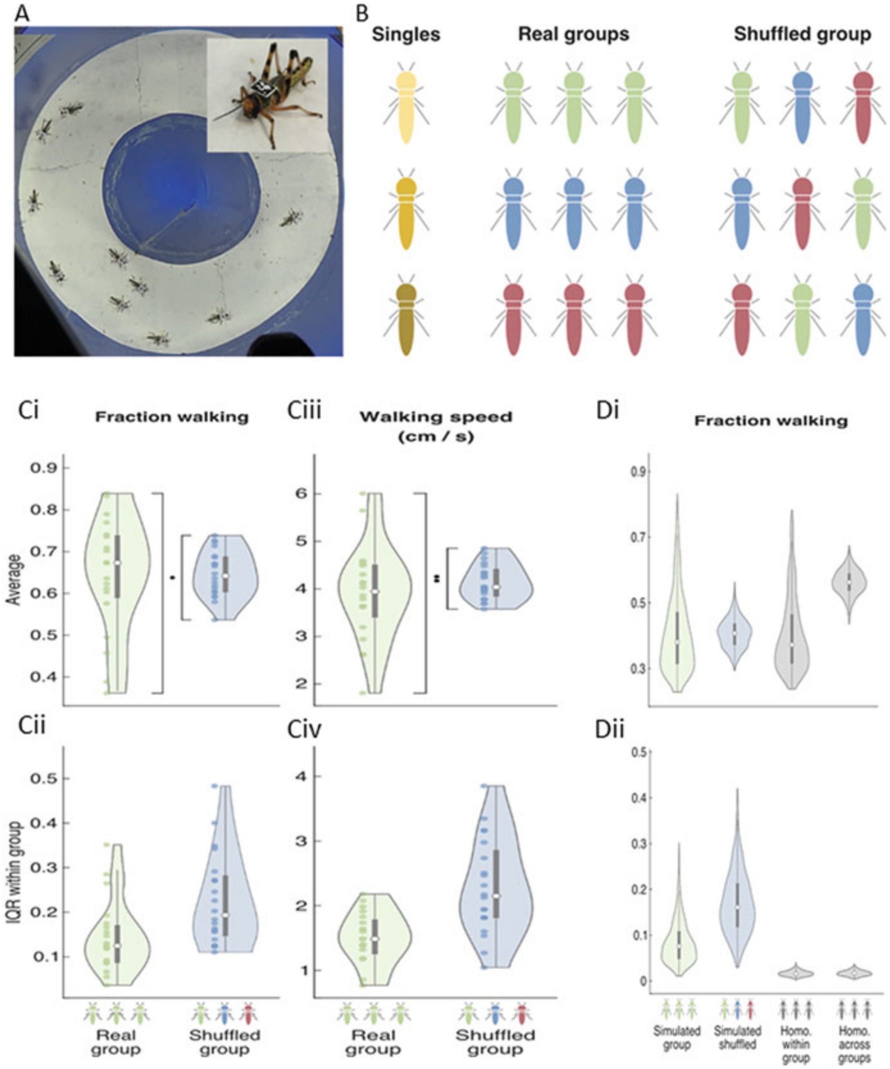


Fig. 1 Heterogeneity of locust swarms. Experiments with marching locust in a circular arena showed that locust groups developed unique, group-specific behavioral characteristics, reflected in large intergroup and small intragroup variance. (**a**) Picture of the experimental setup. (**b**) Data comprised three types: single animals in the arena, groups of 10 animals in the arena (real groups), and fictive groups constructed by shuffling the data of the real groups (shuffled groups). (**c**) Example kinematic results showing the median (ci) and inter quartile range (IQR) among the groups' members (cii) in the fraction of time spent walking, the median (ciii) and IQR (civ) in walking speeds. While different groups show different kinematic properties, within each group (i.e., among the individuals) the differences are significantly lower. This indicates that each group develops a distinctive internal dynamic with specific kinematic features, which is, on one side unique to the group, and on the other side, practiced by all group members similarly. (**d**) Results from a simplified Markov-chain model with parameters that were either derived from experiments with real groups, the shuffled groups or homogeneous ones (same for all group members) equal to the average value of each simulated group (homogenized within groups), or the average of all simulated groups (homogenized across groups). (di) The median in each group. (dii) The IQR. Reproduced from Knebel et al. (2019)

exhibited in control conditions, before and during the introduction to the group. These findings (see Fig. 2) suggest the existence of a distinct behavioral mode in the individual termed a “collective-motion-state.” This state is long lasting, not induced by crowding per se, but only by experiencing collective motion, and characterized by behavioral adaptation to a social context. It was shown that the “collective-motion-state” improves the group’s ability to maintain inter-individual order and proximity. Simulations verify that this behavioral state shortens the average time an animal rejoins the swarm if it departs from it (Knebel et al. 2021). Thus, different socio-environmental circumstances and experiences shape the behaviors of individuals to fit and strengthen the structure of the collective behavior.

The existence of such a “collective-motion-state” is an extreme example of adaptable interactions that enhance a swarms’ stability. It suggests that collective motion is not only an emergent property of the group, but is also dependent on a behavioral mode, rooted in endogenous mechanisms of individuals.

3.4 Microorganisms

Self-organizing emergent phenomena bear critical biological consequences on bacterial colonies and their ability to expand and survive (Grafke et al. 2017; Zuo and Wu 2020). Hence, the properties of mixed swarms are of great significance to our understanding of realistic bacterial colonies.

We begin with a macroscopic point of view that studies the effect of mixed bacterial populations on the overall structure and expansion rate of the entire colony. To this end, we present here new experiments with mixed *B. subtilis* mutants with different cell lengths.[1] Figure 3 shows results obtained with mixed colonies of wild-type and one of three other mutants that vary in their mean length (but have the same width). To distinguish between the strains in a colony, the strains were labeled with a green or red fluorescence protein. Figure 3a is a global view of the colony, showing qualitatively the spatial distribution of the different strains (WT and long mutant) approximately 5 h after inoculation. Figure 3b shows the fraction (in terms of the surface coverage) of the mutant strain at the tip of the expanding colony. On their own, the colony’s expansion rate is independent of the cell shape, regardless of

[1] Experimental conditions: Rapidly/slow moving colonies were grown on soft (0.5%) or hard agar (0.9%) plates, respectively, supplemented with 2 g/l peptone. These growth conditions certify the same expansion rates for all strains while grown separately. Strains used are: “short” DS1470 with aspect ratio 4.1 ± 1.4, “medium” DS860 with aspect ratio 4.7 ± 0.8, wild-type (also medium length) with aspect ratio 4.9 ± 1.7, and “long” DS858 with aspect ratio 8.0 ± 2.3. This method of fluorescence labelling does not affect cell motility, surfactant production, colonial expansion speed or any other quantity that we have tested. The growing colonies were incubated at 30 °C and 95% RH, developed a quasi-circular colonial pattern and were examined microscopically to obtain the ratio between strains at the colonial edge of (Zeiss Axio Imager Z2 at 40×, NEO Andor, Optosplit II). Initially, all the strains were tested axenically for their expansion colonial speed, yielding a fair similarity between them all.

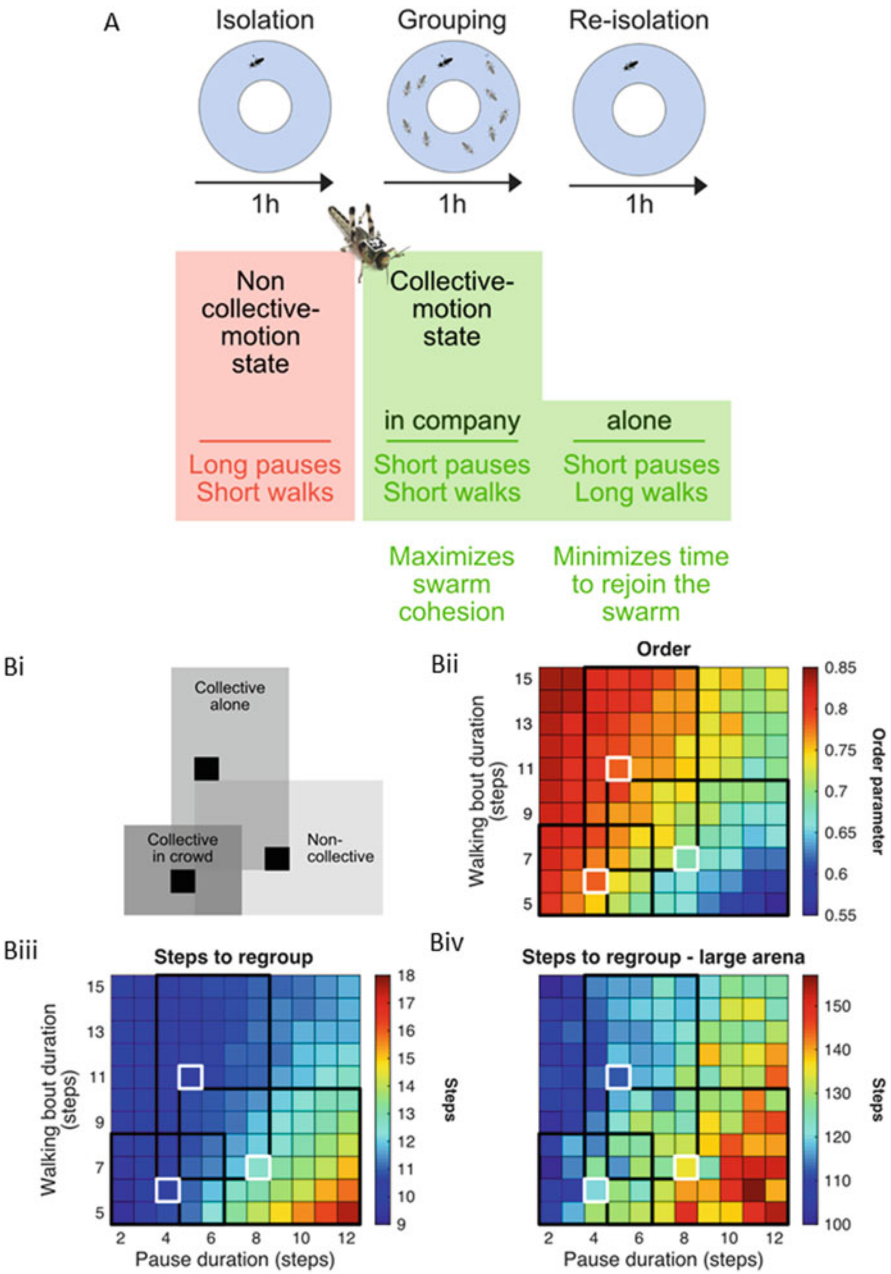


Fig. 2 Collective motion as a distinct behavioral state of the individual. (**a**) A schematic flow of the experimental procedure. The experiments comprised the following consecutive stages: (1) isolation for 1 h in the arena; (2) grouping for 1 h; and (3) re-isolation for 1 h. Each stage is characterized by a different internal state with unique kinematic characteristics. In particular, in each stage animals show different average walking bout and pause durations. (**b**) Agent-based simulations show the influence of different walking bout and pause durations on the collectivity parameters in simulated swarms. (bi) The regions in parameters space indicating the behavioral states. (bii) The order parameter (norm of the average), (biii) The average number of steps to regroup in a small arena (comparable to experimental conditions) and in a larger one (biv). Simulations show that these states may be advantageous for the swarm integrity, shortening the regrouping time if an animal gets separated from the swarm. Reproduced from (Knebel et al. 2021)

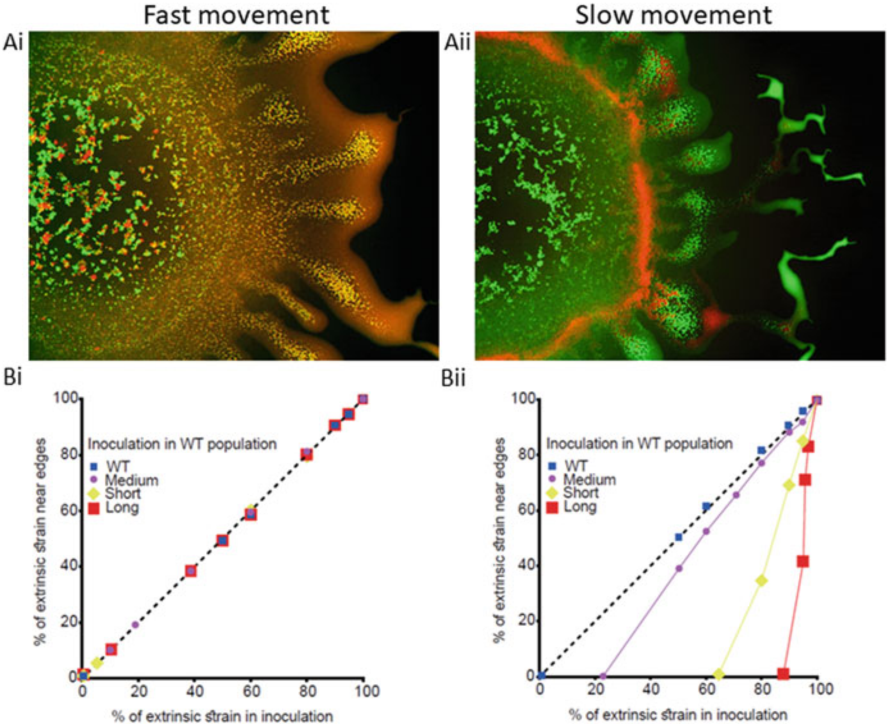


Fig. 3 The macroscopic density distribution in a bacterial colony of *Bacillus subtilis* with a mixed population of cells with different lengths. (**a**) Fluorescent microscope image showing wild-type cells (∼7 μm length, green) and an extra-long mutant (∼19 μm length, red). The scale window size is about 1 cm. (**b**) The fraction of the mutant strain at the edge of the colony as a function of the fraction at the inoculum. On a soft, moist subtract, all cell types can move easily. As a result, the two populations are well-mixed on the macroscopic scale and both populations make it to the front of the colony, where nutrients are abundant. On a hard, dry subtract, wild-type cells spread faster compared to other mutants (either too short or too long) and the colony segregates into wild-type-rich and mutant-rich regions. Thus, the details of the dynamics and the interaction between the species and the environment determine the macroscopic state of the swarm

the substrate on which they are grown. However, in mixed colonies, results depend on the hardness of the substrate. On soft agar (left column), all bacteria can move easily and fast. As a result, the ratios between strains in the initial inoculum are same as the one obtained at the colony's edge, indicating that all strains migrate at similar speeds with no apparent competition between them. On the other hand, movement on hard agar (right column) is slower, and the ratio between strains in the initial inoculum is different from the one obtained at the edge. In particular, strains that are either shorter or longer compared to the wild-type show a disadvantage. For example, in a mix of short cells and wild-type, the short cells do not make it to the edge at all unless their initial concentration is above 65%.

In a second example, Deforet et al. (Deforet et al. 2019) studied a mixed colony of wild-type *Pseudomonas aeruginosa* with a mutant that disperses $\sim 100\%$ faster but grows $\sim 10\%$ slower (possibly due to resources redirected to grow extra flagellum). Thus, their experiment tests the trade-off between growth and dispersal. Although a model predicts that in some cases, better growth rate may win over faster dispersal, all experiments showed the opposite, i.e., that getting first to the front of the colony (where nutrients are abundant) is the bottleneck for fast colony expansion.

The two experiments described above clearly show that the coupling between species competition and the environment results in complex macroscopic spatial patterns that is difficult to predict based on first principles.

Next, we concentrate on the microscopic properties of swarming bacteria. One of the main challenges in studying heterogeneous systems of microorganisms is in distinguishing between biological and physical interactions. Microorganisms belonging to different species and strains typically have many differences, ranging from mechanical properties such as cell size, different physical responses to external ques. (e.g., different effective drift-diffusion parameters) to species-specific metabolic processes. In order to untangle all these effects, Peled et al. (2021) focused on mixtures of same species swarms differing only in cell size.

Bacterial swarms are composed of millions of flagellated, self-propelled cells that move coherently in dynamic clusters forming whirls and jets. Dominated by hydrodynamic interactions and cell–cell steric forces, the characteristics of the individuals dictate the dynamics of the group (Be'er and Ariel 2019). Empirically, active cells tend to elongate prior to swarming, and their length (or rather aspect ratio) was shown to play a crucial role in determining their collective statistics, suggesting a length selection mechanism. In Ilkanaiv et al. (2017), it was shown that although homogeneous colonies of bacteria with different aspect ratios spread at the same speed, their microscopic motion differs significantly. Both short and long strains were moving slower, exhibiting non-Gaussian statistics; however, the wild-type, and strains that are close in size to the wild-type, were moving faster with Gaussian statistics. Overall, bacteria are thought to have adapted their physics to optimize the principle functions assumed for efficient swarming.

Surprisingly, introducing a small number of cells with a different length than the majority can have a significant effect on the dynamics of the swarm (Peled et al. 2021). The cooperative action of many short cells mixed with a few longer cells leads to longer spatial correlations (indicating a more ordered swarming pattern) and higher average cell speeds. Figures 4 and 5 show that a small number of long cells helps organizing the dynamics of the bacterial colony, with long cells acting as nucleation sites, around which aggregates of short, rapidly moving cells can form. Increasing the fraction of long cells (i.e., increasing heterogeneity), the average speed drops as the long cells become jammed, serving as a bottleneck for efficient swarming. The impact of long cells was reproduced in a simple model based on hydrodynamic interactions, indicating a purely physical mechanism behind the beneficial effects of a few long cells on spatial organization and motion of all cells in the swarm. To the best of our knowledge, this is a first example showing that heterogeneity can promote order and increase swarm speeds.

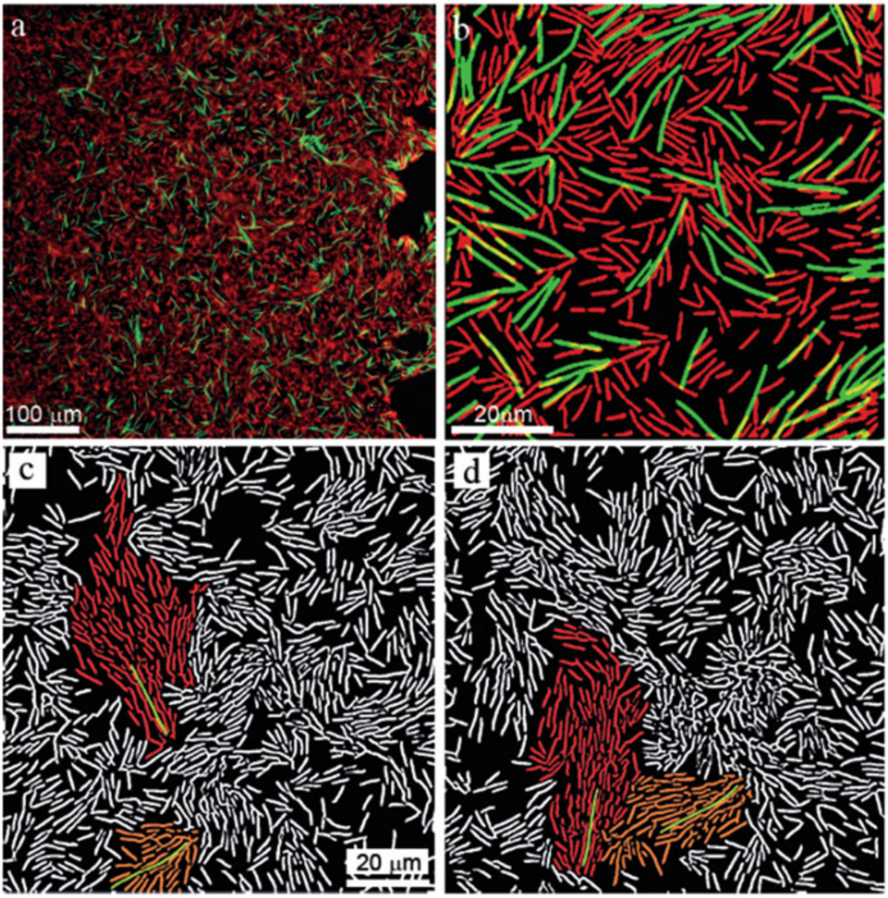


Fig. 4 The microscopic density distribution in a bacterial colony of *Bacillus subtilis* with a mixed population of cells with different lengths. (**a**, **b**) The wild-type in red and the elongated cells in green. When the mixing ratio is about 50:50, the swarm is well-mixed. (**c**, **d**) Cohesive moving clusters (false-colored for illustration purposes). When the fraction of long cells is small, short cells cluster around elongated ones, moving together. The figures shows two such clusters in two consecutive snapshots, 0.3 s apart. We find that the ratio between the two populations determines the spatial distribution and the dynamics of the swarm. Reproduced from Peled et al. (2021)

4 Modeling Heterogeneous Collective Motion

Theory and simulations of active matter establish that heterogeneous systems of self-propelled agents show a range of interesting dynamics and a wealth of unique phases that depend on the properties of individuals. In accordance with the discussions above, researchers studied two main sources of variability. The first type assumes fixed properties (at least on the time scale of the dynamics of interest), for example, individuals with different velocities (Hemelrijk and Hildenbrandt 2011;

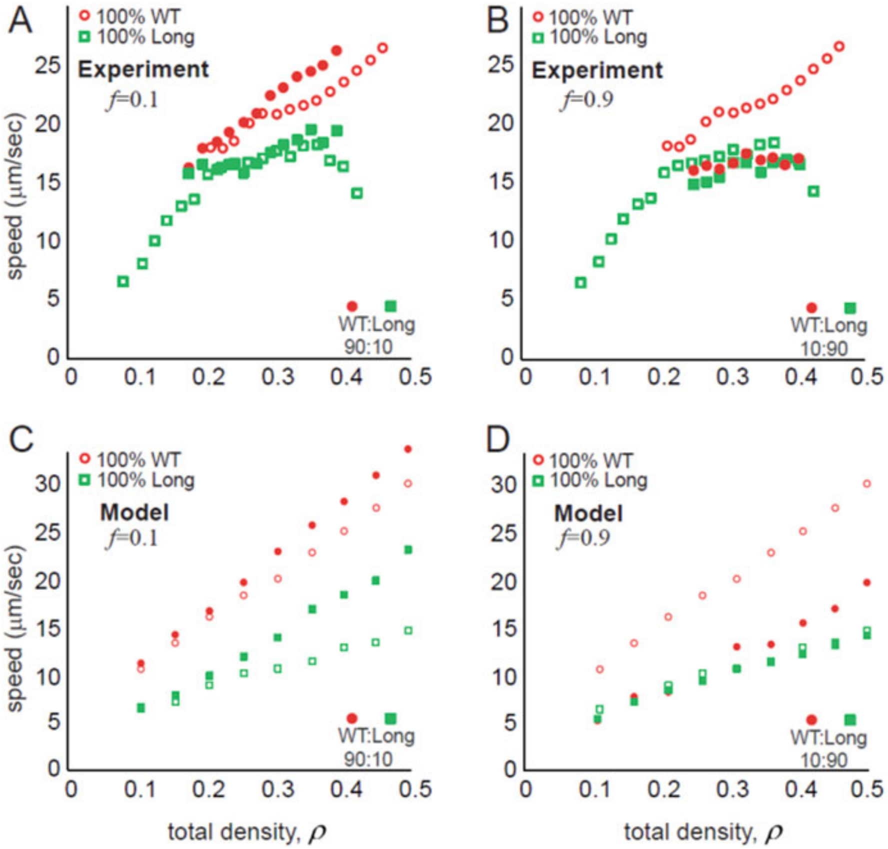


Fig. 5 Microscopic dynamics in a mixed bacterial swarm. (**a**, **b**) Experiments and (**c**, **d**) Agent-based simulations. (**a**, **c**) The average swarm speeds in heterogeneous populations with a small fraction of elongated cells (10%) as a function of the total area fraction. (**b**, **d**). The average swarm speeds in heterogeneous populations with a large fraction of elongated cells (90%). Surprisingly, introducing a small number of cells with a different aspect ratio than the majority increases swarm speeds. Reproduced from Peled et al. (2021)

McCandlish et al. 2012; Schweitzer and Schimansky-Geier 1994; Singh and Mishra 2020), noise sensitivity (Ariel et al. 2015; Benisty et al. 2015; Menzel 2012; Netzer et al. 2019), sensitivity to external cues (Book et al. 2017), and particle-to-particle interactions (Bera and Sood 2020; Copenhagen et al. 2016; Hemelrijk and Kunz 2005; Khodygo et al. 2019). It was found that the effect of heterogeneity ranges from trivial (the mixed system is equivalent to an average homogeneous one) to singular (one of the sub-populations dominates the dynamics of the group as a whole) (Ariel et al. 2015). The second type of heterogeneity refers to identical individuals whose properties change due to different local environments, for example, local density of conspecifics (Castellano et al. 2009; Cates and Tailleur 2015; Helbing 2001) or topology of the environment (Berdahl et al. 2013; Khodygo et al. 2019; Shklarsh et al. 2011; Torney et al. 2009). Such differences may have a significant effect on the ability of swarms to organize and, in particular, navigate towards required goals (Berdahl et al. 2013; Khodygo et al. 2019; Shklarsh et al. 2011; Torney et al. 2009). Coupling between the different populations and heterogeneous environments may lead to the evolution of territories (Alsenafi and Barbaro 2021). Note that here, we do not consider uniform distribution of obstacles (e.g., Chepizhko et al. 2013; Rahmani et al. 2021) as heterogeneity.

4.1 Continuous Models

To the best of our knowledge, the first theoretical works on the effect of heterogeneity on collective motion approached the question from the point of view of population dynamics in heterogeneous environments. In the mid-1970s, Comins and Blatt (1974), Roff (1974a, b), Roughgarden (1974) and subsequently Levin (1976) studied the dynamics of a finite number of migrating populations using continuous models. Local dynamics was modeled, for example, by logistic growth, while migration was taken into account by a linear diffusion (or a discrete analogue). The biological motivation for this point of view is migration in a patchy environment. Patches (or niches) were characterized by different parameters, for example, carrying capacities. One of the main conclusions of these studies was that movement can have a stabilizing effect on the dynamics, for example, suppressing oscillations in predator–prey models (Comins and Blatt 1974).

At the same time, Horn and MacArthur (1972), followed by Segel and Levin (1976), and Gopalsamy (s1977) considered continuous two-species spatial models. Again, movement was modeled using diffusion. The main goal was to study the effect of migration on the stability of communities. Conditions, in which initially mixed species evolve into spatially segregated regions were found particularly interesting. See Kareiva (1990) for a review of this perspective. The coupling between two-species competition and heterogeneous environments was studied by Dubois, motivated by plankton populations (Dubois 1975) and McMurtrie (1978). They study different forms of models with non-uniform dispersal and drift. For example, McMurtrie (1978) propose a one-dimensional (1D) model involving a two-species predator–prey system of the form

$$
\begin{aligned}
\frac{\partial n}{\partial t} &= an\,(1-bp) + \mu\frac{\partial^2 n}{\partial x^2} + \alpha\,\mathrm{sgn}(x)\frac{\partial n}{\partial x}\\
\frac{\partial p}{\partial t} &= dp\,(-1+cn) + \nu\frac{\partial^2 p}{\partial x^2} + \beta\,\mathrm{sgn}(x)\frac{\partial p}{\partial x},
\end{aligned}
$$

where $n(t,x)$ and $p(t,x)$ are the density of predators and prey, respectively. The constants a, b, c, and d are the standard Lotka–Volterra parameters, μ and ν are diffusion constants. The terms involving α and β describe preferential dispersal towards the center of the habitat.

Motivated by works on single species spatial distribution patterns, as well as analogue models of gas flow in porous medium, (in the 1980s) modeling shifted towards nonlinear reaction-diffusion equations in which the flux term depends on the local concentration (Aronson 1980; Namba 1980, 1989). Thus, the description inherently takes into account a heterogeneous environment, as movement is density dependent. Two-species versions were also explored (Bertsch et al. 1984; Mimura and Kawasaki 1980; Namba and Mimura 1980; Shigesada et al. 1979; Witelski 1997). The main goals were again to classify under which conditions populations mix or segregate. For example, Mimura and Kawasaki (1980) study a 1D predator–prey model with nonlinear self and cross-diffusion of the form

$$\begin{aligned}\frac{\partial n}{\partial t} &= an\left(1-bp-en\right)+\frac{\partial^2}{\partial x^2}\left[\left(\alpha+\beta_1 p\right)n\right]\\ \frac{\partial p}{\partial t} &= dp\left(-1+cn+fp\right)+\frac{\partial^2}{\partial x^2}\left[\left(\alpha+\beta_1 n\right)p\right].\end{aligned}$$

Traveling wave solutions (Gurtin and Pipkin 1984) later (in the 1990s) proved to be important to modeling of expanding bacterial colonies (Ben-Jacob et al. 2000).

New forms of models were derived by coarse-graining agent-based models. The first approaches, such as those of Toner-Tu (Toner et al. 2005) and Swift-Hohenberg (Wensink et al. 2012) applied phenomenological models that were based on physical principles and the underlying symmetries in collective systems. More rigorous approaches derived coarse-grained equations of agent-based models under appropriate limits (e.g., Carrillo et al. 2010; Degond and Motsch 2008; Ha and Tadmor 2008). Much research involves density dependent parameters (Frouvelle 2012), in particular speed dependence (see Cates and Tailleur 2015; Degond et al. 2017) and the references therein. For example, derived from microscopic considerations, Cates and Tailleur (2015) consider the stochastic partial differential equation

$$\begin{aligned}\frac{\partial \rho\left(t,x\right)}{\partial t} &= -\nabla\cdot J\\ J &= -D\left(\rho\right)\nabla\rho+V\left(\rho\right)\rho+\sqrt{2D\left(\rho\right)\rho}\,\dot{W},\end{aligned}$$

where $\dot{W}$ is white noise and $V(\rho)$, $D(\rho)$ are density dependent drift-diffusion coefficients, typically taken as linearly decreasing in ρ. The main conclusion is that density dependent motility may lead to the so-called motility induced phase separation, in which the system self-segregates into coexisting low-density/high-speed that are characterized by high-density/low-speed regions. Density dependent speeds are also fundamental to understanding traffic and pedestrian dynamics through the so-called fundamental diagram, relating the flux of individuals to the local density (Castellano et al. 2009; Helbing 2001).

Similarly, considerable research has been devoted to continuous descriptions of binary self-propelled particle mixtures. Models included variability in motility (Book et al. 2017; Deforet et al. 2019; Navoret 2013), noise (Menzel 2012), strength of alignment (Yllanes et al. 2017), cross interactions between species (Burger et al. 2018; Chertock et al. 2019; Di Francesco and Fagioli 2013), or cross-diffusion

(Alsenafi and Barbaro 2021; Book et al. 2017; Carrillo et al. 2018, 2020; Di Francesco et al. 2018). The results vary significantly in the level or rigor. Again, a key question of interest is the effect of heterogeneity on the order-disorder transition and spatial phase segregation. For example, Carrillo et al. (2018) extend previous models and study a nonlinear and non-local model with cross-diffusion of the form

$$\begin{aligned} \frac{\partial n}{\partial t} &= \nabla \cdot \left[n \nabla \left(W_{11} * n + W_{12} * p + \epsilon \left(n + p \right) \right) \right] \\ \frac{\partial p}{\partial t} &= \nabla \cdot \left[p \nabla \left(W_{22} * p + W_{21} * n + \epsilon \left(n + p \right) \right) \right], \end{aligned}$$

where W_{ij} are interaction terms, typically of power-law form (e.g., Lennard-Jones), characteristic functions (steric repulsion), or exponential (Morse potential), $\epsilon > 0$ is the coefficient of cross-diffusion (Carrillo et al. 2018), and * denotes the convolution operator.

4.2 Agent-Based Models

With the availability of large-scale computer simulations and the success of newly suggested simplified ABMs (Giardina 2008; Vicsek and Zafeiris 2012), much of the theoretical research on collective motion, especially models that study concrete biological systems, shifted towards discrete models. Most works are either based on the three zones model of Aoki and Reynolds (Aoki 1982; Reynolds 1987) or the Vicsek model (Vicsek et al. 1995). In the three zones model, each agent is either repelled, aligned, or attracted to conspecifics with fixed interaction ranges. Typically, the repulsion range, which describes collision avoidance, is shortest. The attraction range, allowing long range group cohesion, is the largest. In between, agents align their direction of movements according to the local average. In contrast, the Vicsek model only has local alignment, which is countered by added angular noise. To be precise, N particles with positions $x_i \in \mathbb{R}^2$ and velocity $v_i \in \mathbb{R}^2$ move with a fixed speed $|v_i| = v_0$ in a 2D rectangular domain with periodic boundaries. At each simulation step, each agent aligns with the average direction of movement of all particles within a fixed interaction range. Then, the average direction is perturbed randomly. The equations of motion at each simulation step are

$$\begin{aligned} x_i &\leftarrow x_i + v_i \\ v_i &\leftarrow \Theta_{\phi_i} \frac{\sum_{\{j:|x_i - x_j| \le R\}} v_i}{\left| \sum_{\{j:|x_i - x_j| \le R\}} v_i \right|} v_0 , \end{aligned}$$

where ϕ_i are independent random variables, uniformly distributed in the segment $[-\sigma\pi, \sigma\pi]$, $0 \le \sigma \le 1$. The main prediction of this model is the characterization of two regimes (or phases), depending on the noise level σ and the average density—a disordered phase, in which the average velocity agents goes to zero in the limit of an infinite system, and an ordered phase in which it does not (Vicsek et al. 1995).

Context-dependent interactions within the three-zone model were first studied by Torney et al. (2009). The main idea was that individuals weigh their own information regarding the environment and the local movement of conspecifics dynamically, according to local conditions or available information. In Shklarsh et al. (2011), a particular simple adaptable 2D model studied the rate in which a collection of SPPs can reach a maximum of a fixed external potential $c(x)$. The model, which is essentially a three zones model, assumes that in each simulated step, the direction in which an agent moves, denoted $\hat{d}_i$, is a weighted sum of two terms: $\hat{u}_i$, denoting group interaction following the three zones model, and $\hat{v}_i$, which is the particle velocity at the previous step

$$\begin{array}{l} \overline{d}_i \leftarrow \hat{u}_i + w\hat{v}_i \\ \hat{d}_i = \overline{d}_i / \left|\overline{d}_i\right| \end{array}.$$

The main idea of Shklarsh et al. (2011) is to make the weight w a function of the environment $c(x)$. Denoting by $\Delta c_i(t)$ the difference in $c(x)$ between two consecutive steps of agent i, they take

$$w_i(t) = \begin{cases} 1 & |\Delta c_i(t)| > \text{const} \\ 0 & \text{otherwise} \end{cases}.$$

In words, the external cue shuts down if the gradient in $c(x)$ in the direction of movement is too small. This strategy proved efficient in sensing the environment (Berdahl et al. 2013; Shaukat and Chitre 2016). Other examples of adaptable models, e.g., density dependent speeds (Mishra et al. 2012), were found to be sufficient to induce phase separation between dense and dilute fluid phases (Cates and Tailleur 2015) and to increase the stability of swarms (Gorbonos and Gov 2017; Ling et al. 2019).

Over the past decade, collective motion of binary self-propelled particle mixtures has been extensively researched theoretically using agent-based models. The effect of fixed variability (i.e., non-adaptable) in motility (Agrawal and Babu 2018; Benisty et al. 2015; Copenhagen et al. 2016; Khodygo et al. 2019; Kumar et al. 2014; McCandlish et al. 2012), weight of alignment interactions (del Mar Delgado et al. 2018; Knebel et al. 2021; Kunz and Hemelrijk 2003; Peled et al. 2021; Soni et al. 2020), effective noise (Ariel et al. 2015; Guisandez et al. 2017; Menzel 2012), and interaction range (Farine et al. 2017) were studied. The main questions considered are the effect of heterogeneity on the ability of swarms to form ordered phases (Agrawal and Babu 2018; Ariel et al. 2015; Benisty et al. 2015; Copenhagen et al. 2016; del Mar Delgado et al. 2018; Kumar et al. 2014; Peled et al. 2021; Soni et al. 2020), the type of the order-disorder transition (first or second order) (Guisandez et al. 2017), spatial segregation of the two species (Copenhagen et al. 2016; Khodygo et al. 2019; McCandlish et al. 2012), or the organization within the swarm (Farine et al. 2017; Hemelrijk and Hildenbrandt 2008, 2011; Hemelrijk and Kunz 2005; Peled et al. 2021), and the rate of convergence towards the ordered steady state (Knebel et al. 2021). Not surprisingly, and as confirmed experimentally (see the previous section), heterogeneity typically lowers order. If the variation between individuals is sufficiently large, the ordered phase may either disappear completely or, alternatively, the system may segregate into coexisting, spatially separated phases. As mentioned before, low order has its own benefits—and perhaps reducing order by heterogeneity is not a bug but a feature.

For example, Ariel et al. (2015) study, using simulations, a variation of the Vicsek model with two populations that are distinguished by the amount of noise they have. In the original Vicsek model, the noise level σ is the same for all particles. In Ariel et al. (2015), it is assumed that a fraction f of the agents has noise level σ_1, while the rest have σ_2. In order to quantitatively compare homogeneous and heterogeneous systems, one needs to identify the appropriate statistics (corresponding to the relevant thermodynamic variables). Following Porfiri and Ariel (2016) and Ariel et al. (2015), the circular mean of distribution of the random turns plays the role of (1 minus) an effective temperature, in the sense that it determines the order parameter and phase (here, an ordered phase means that the mode in the distribution of the instantaneous order parameter is not zero). Moreover, it satisfies a fluctuation–dissipation relation (Porfiri and Ariel 2016). In heterogeneous systems, the two sub-populations interact non-additively: Within a large range of parameters, the dynamics of the system can be described by an equivalent homogeneous one with the same average temperature $fT_1+(1\text{-}f)\,T_2$. However, if one of the sub-populations is sufficiently "cold," i.e., σ_1 or σ_2 (or equivalentrly, the effective temperaturese T_1 or T_2) is below a threshold, it dominates the dynamics of the group as a whole. Specifically, it determines the phase and order parameter of the mixed system, see Fig. 6 for a phase diagram. Interestingly, this phenomenon does not occur in mean-field random-network models of collective motion, but depends on emergence of spatial heterogeneities (Netzer et al. 2019).

Finally, ABMs were used to study the effect of a social structure of the ability of swarms to synchronize. Leadership was studied in Couzin et al. (2005) and Garland et al. (2018). In Xue et al. (2020), a hierarchical swarm model in the spirit of the Vicsek model showed that introducing a simple hierarchical structure (via a linear ordering of agents) not only shifts the order–disorder phase transition, but also changes its type (first or second order).

4.3 Specific Examples: Locust

A few groups attempted to address the dynamics of locust swarms theoretically. Topaz et al. (2012) studied a continuous binary-system model, describing the density of solitarious and gregarious locusts. The main assumption is that individuals can switch between phases (solitarious and gregarious) with rates that depend on the overall local density. Thus, while the sum of the two densities is a conserved quantity, satisfying a continuity equation, each density on its own does not. The difference between the phases is in its interaction with conspecifics: while

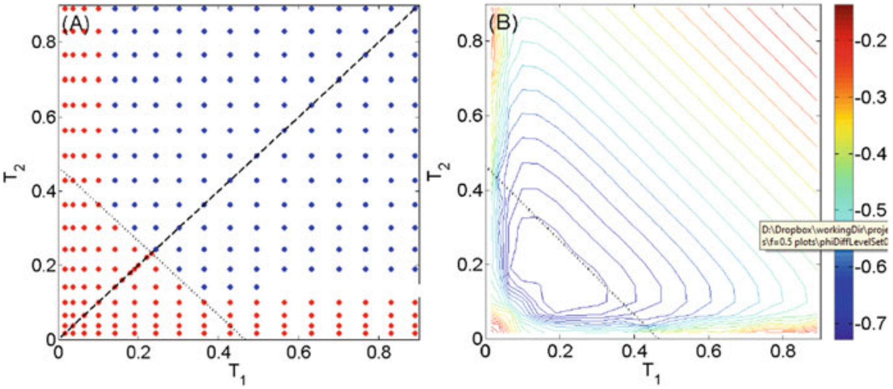


Fig. 6 Simulation results for a two-species Vicsek model with distinct noise levels. A heterogeneous system with 50,000 particles, half with effective temperature (the circular mean of random turns) T_1, and half with effective temperature T_2. (**a**) The phase diagram. Red dots indicate an ordered phase, while blue dots are disordered. (**b**) The difference between the observed temperature (1-order parameter) and the average effective temperature $(T_1 + T_2)/2$. If T_1 or T_2 is small enough, level curves are close to horizontal or vertical lines. Otherwise, they are diagonal lines, indicating a constant temperature. The dashed curve shows the homogeneous $T_1 = T_2$ line. The dotted line is the constant temperature curve passing through the homogeneous critical temperature. Reproduced from Ariel et al. (2015)

solitarious individuals are repelled from other locusts, gregarious individuals are attracted. The interaction term is non-local. The model is used to study band formation. In particular, numerical solutions reveal transiently traveling clumps of gregarious insects.

Another binary-system model, taking only gregarious locust into account, studied the impact of the pause-and-go walking pattern of locust on the spatial distribution of marching bands. In Bernoff et al. (2020), the authors study both an ABM and a simple continuous realization of a two-species model describing stationary and moving insects. Heterogeneous environments are also taken into account in the form of position dependent resource consumption rate. One of the main new assumptions is that the rate at which locusts transition between moving and stationary (and vice versa) is enhanced (diminished) by resource abundance.

Lastly, a recent work (Georgiou et al. 2020) combines the two approaches, studying the dynamics of solitarious and gregarious insects in a heterogeneous environment in terms of the available food resources.

4.4 Specific Examples: Microorganisms and Cells

Previous modeling approaches of heterogeneous active matter or self-propelled particles have been used, with some levels of success to study several aspects of mixed bacterial communities (Blanchard and Lu 2015; Book et al. 2017; Deforet et al. 2019; Kai and Piechulla 2018; Kumar et al. 2014). For example, on the macroscopic, colony-wide scale, continuous models of mixed bacterial colonies with different motility and growth rates show the balance between reproduction rates and the importance of moving towards the colony edge, where nutrients are abundant (Book et al. 2017; Deforet et al. 2019). Peled et al. (2021) studied a two-species agent-based model (with different cell-length) that is derived from the balance of forces and torques on each cell. The model follows the approach of Ariel et al. (2018) and Ryan et al. (2011), assuming each bacterium is essentially a point dipole where the size is incorporated through an excluded-volume potential and the shape is accounted for in the interaction of the point dipole's orientation with the fluid. The model reproduces the speed dependence of both cell types at the entire range of densities tested. However, in contrast with experiments, the simulated spatial distributions of short (wild-type) and long cells are not correlated. Therefore, hydrodynamic models of swarming bacteria fall short at describing the full breadth of the dynamics.

A detailed, mixed population, agent-based 2D model that includes both excluded volume and hydrodynamic interactions was studied by Jeckel et al. (2019). In this model, agents are elongated ellipsoids with a distribution of lengths, motility, and friction coefficients, as observed experimentally for different phases during the growth of bacterial colonies. The model successfully reproduces the motile phases observed experimentally in an expanding colony of swarming *B. subtilis*.

Finally, we briefly discuss collective cell migration, which plays a pivotal role in a range of biological processes such as wound healing, cancer invasion and development (Schumacher et al. 2016, 2017). Heterogeneity, both between cells and in the environment (typically non-uniform tissues) has been identified as a key parameter in the regulation and differentiation of cells, for example, in development of tips vs. stalks (Rørth 2012). Of course, cells are not organisms. However, the theory of collective cell migration shares many of the universal properties of other collective motion phenomena (Chauviere et al. 2007; Gavagnin and Yates 2018; Szabo et al. 2006), for example, a kinetic phase transition from a disordered to ordered state (Szabo et al. 2006), spatial segregation and "task specification" (Rørth 2012).

5 Summary and Concluding Remarks

Variability is inherent to practically all groups of organisms. As discussed above, the sources of variations among members of a group are diverse, from differences rooted in ontogeny and development, via changes due to physiological adaptations, to distinct behavioral states. Accordingly, the variations may be transient or lasting. Collective motion, manifested by coordinated or synchronized group movement requires, by definition, a level of similarity between the individuals composing the group. Furthermore, the groups exert a homogenizing effects on its members.

This alleged discrepancy, or tug-of-war type interaction, between the group and the individual (i.e., variability vs. homogeneity) is at the basis of much of the rich and complex dynamics seen in collective motion.

This review presents both the state-of-the-art and a historical perspective of experimental and theoretical aspects of heterogeneity in real, natural swarms. We focus on natural-biological systems only; however, the main points are also relevant to humans and human made systems (pedestrians, cars, robots, etc.). The conclusion of most theoretical work is rather straightforward, i.e., a higher heterogeneity diminishes the order, as one could intuitively expect. However, as evident from the different examples discussed, heterogeneity may be contributory and even instrumental in the interaction leading to the self-emergence of collective motion. The disparity between the rather simplistic theoretical conclusions and the known biological prevalence and significance of variability in nature raises a major open question (critically important to biological systems) of the ecological and evolutionary consequences of heterogeneity within collectives. In particular, it is not clear under what circumstances is heterogeneity, and its consequences on collective motion, evolutionary advantageous, or is it merely a natural, unavoidable reality that interferes with collectivity. Such considerations are often not taken into account in simplified mathematical models.

The challenges ahead of us include deciphering these interactions in new, diverse systems and at different types of environments. Also, there is currently very little work on continuous distribution of heterogeneities, as well as on coupling between different properties, which are more biologically realistic. By utilizing a comparative approach for developing general rules, we will be able to provide a further solid theoretical framework for the development of collectivity in light of variability and heterogeneity.

Acknowledgements This research has been supported by the Israel Science Foundation, research grants 373/16 and 2306/18.

References

1. Agrawal, A., Babu, S.B., 2018. Self-organization in a bimotility mixture of model microswimmers. Phys. Rev. E 97, 20401.
2. Ákos, Z., Beck, R., Nagy, M., Vicsek, T., Kubinyi, E., 2014. Leadership and path characteristics during walks are linked to dominance order and individual traits in dogs. PLoS Comput Biol 10, e1003446.
3. Alsenafi, A., Barbaro, A.B.T., 2021. A Multispecies Cross-Diffusion Model for Territorial Development. Mathematics 9, 1428.
4. Amichay, G., Ariel, G., Ayali, A., 2016. The effect of changing topography on the coordinated marching of locust nymphs. PeerJ 4, e2742.
5. Aoki, I., 1982. A simulation study on the schooling mechanism in fish. Bull. Japanese Soc. Sci. Fish.
6. Ariel, G., Ayali, A., 2015. Locust collective motion and its modeling. PLoS Comput. Biol. 11, e1004522.

7. Ariel, G., Ophir, Y., Levi, S., Ben-Jacob, E., Ayali, A., 2014. Individual pause-and-go motion is instrumental to the formation and maintenance of swarms of marching locust nymphs. PLoS One 9, e101636.
8. Ariel, G., Rimer, O., Ben-Jacob, E., 2015. Order–disorder phase transition in heterogeneous populations of self-propelled particles. J. Stat. Phys. 158, 579–588.
9. Ariel, G., Sidortsov, M., Ryan, S.D., Heidenreich, S., Bär, M., Be'er, A., 2018. Collective dynamics of two-dimensional swimming bacteria: Experiments and models. Phys. Rev. E 98, 32415.
10. Aronson, D.G., 1980. Density-dependent interaction–diffusion systems, in: Dynamics and Modelling of Reactive Systems. Elsevier, pp. 161–176.
11. Bär, M., Großmann, R., Heidenreich, S., Peruani, F., 2020. Self-propelled rods: Insights and perspectives for active matter. Annu. Rev. Condens. Matter Phys. 11, 441–466.
12. Barber, I., Ruxton, G.D., 2000. The importance of stable schooling: do familiar sticklebacks stick together? Proc. R. Soc. London. Ser. B Biol. Sci. 267, 151–155.
13. Barber, I., Wright, H.A., 2001. How strong are familiarity preferences in shoaling fish? Anim. Behav. 61, 975–979.
14. Barnett, I., Khanna, T., Onnela, J.-P., 2016. Social and spatial clustering of people at humanity's largest gathering. PLoS One 11, e0156794.
15. Bazazi, S., Bartumeus, F., Hale, J.J., Couzin, I.D., 2012. Intermittent motion in desert locusts: behavioural complexity in simple environments. PLoS Comput Biol 8, e1002498.
16. Be'er, A., Ariel, G., 2019. A statistical physics view of swarming bacteria. Mov. Ecol. 7, 1–17.
17. Ben-Jacob, E., Cohen, I., Levine, H., 2000. Cooperative self-organization of microorganisms. Adv. Phys. 49, 395–554.
18. Ben-Jacob, E., Finkelshtein, A., Ariel, G., Ingham, C., 2016. Multispecies swarms of social microorganisms as moving ecosystems. Trends Microbiol. 24, 257–269.
19. Benisty, S., Ben-Jacob, E., Ariel, G., Be'er, A., 2015. Antibiotic-induced anomalous statistics of collective bacterial swarming. Phys. Rev. Lett. 114, 18105.
20. Bera, P.K., Sood, A.K., 2020. Motile dissenters disrupt the flocking of active granular matter. Phys. Rev. E 101, 52615.
21. Berdahl, A., Torney, C.J., Ioannou, C.C., Faria, J.J., Couzin, I.D., 2013. Emergent sensing of complex environments by mobile animal groups. Science (80-.). 339, 574–576.
22. Bernoff, A.J., Culshaw-Maurer, M., Everett, R.A., Hohn, M.E., Strickland, W.C., Weinburd, J., 2020. Agent-based and continuous models of hopper bands for the Australian plague locust: How resource consumption mediates pulse formation and geometry. PLoS Comput. Biol. 16, e1007820.
23. Bertsch, M., Gurtin, M.E., Hilhorst, D., Peletier, L.A., 1984. On interacting populations that disperse to avoid crowding: preservation of segregation. WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER.
24. Blanchard, A.E., Lu, T., 2015. Bacterial social interactions drive the emergence of differential spatial colony structures. BMC Syst. Biol. 9, 1–13.
25. Book, G., Ingham, C., Ariel, G., 2017. Modeling cooperating micro-organisms in antibiotic environment. PLoS One 12, e0190037.
26. Brown, C., Irving, E., 2014. Individual personality traits influence group exploration in a feral guppy population. Behav. Ecol. 25, 95–101.
27. Buhl, J., Sumpter, D.J.T., Couzin, I.D., Hale, J.J., Despland, E., Miller, E.R., Simpson, S.J., 2006. From disorder to order in marching locusts. Science (80-.). 312, 1402–1406.
28. Burger, M., Francesco, M. Di, Fagioli, S., Stevens, A., 2018. Sorting phenomena in a mathematical model for two mutually attracting/repelling species. SIAM J. Math. Anal. 50, 3210–3250.
29. Carrillo, J.A., Filbet, F., Schmidtchen, M., 2020. Convergence of a finite volume scheme for a system of interacting species with cross-diffusion. Numer. Math. 145, 473–511.

30. Carrillo, J.A., Fornasier, M., Toscani, G., Vecil, F., 2010. Particle, kinetic, and hydrodynamic models of swarming, in: Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences. Springer, pp. 297–336.
31. Carrillo, J.A., Huang, Y., Schmidtchen, M., 2018. Zoology of a nonlocal cross-diffusion model for two species. SIAM J. Appl. Math. 78, 1078–1104.
32. Castellano, C., Fortunato, S., Loreto, V., 2009. Statistical physics of social dynamics. Rev. Mod. Phys. 81, 591.
33. Cates, M.E., Tailleur, J., 2015. Motility-induced phase separation. Annu. Rev. Condens. Matter Phys. 6, 219–244.
34. Chauviere, A., Hillen, T., Preziosi, L., 2007. Modeling cell movement in anisotropic and heterogeneous network tissues. Networks Heterog. Media 2, 333.
35. Chepizhko, O., Altmann, E.G., Peruani, F., 2013. Optimal noise maximizes collective motion in heterogeneous media. Phys. Rev. Lett. 110, 238101.
36. Chertock, A., Degond, P., Hecht, S., Vincent, J.-P., 2019. Incompressible limit of a continuum model of tissue growth with segregation for two cell populations [J]. Math. Biosci. Eng. 16, 5804–5835.
37. Comins, H.N., Blatt, D.W.E., 1974. Prey-predator models in spatially heterogeneous environments. J. Theor. Biol. 48, 75–83.
38. Copenhagen, K., Quint, D.A., Gopinathan, A., 2016. Self-organized sorting limits behavioral variability in swarms. Sci. Rep. 6, 1–11.
39. Couzin, I.D., Krause, J., Franks, N.R., Levin, S.A., 2005. Effective leadership and decision-making in animal groups on the move. Nature 433, 513–516.
40. Croft, D.P., James, R., Krause, J., 2008. Exploring animal social networks. Princeton University Press.
41. Deforet, M., Carmona-Fontaine, C., Korolev, K.S., Xavier, J.B., 2019. Evolution at the edge of expanding populations. Am. Nat. 194, 291–305.
42. Degond, P., Henkes, S., Yu, H., 2017. Self-organized hydrodynamics with density-dependent velocity. Kinet. Relat. Model.
43. Degond, P., Motsch, S., 2008. Continuum limit of self-driven particles with orientation interaction. Math. Model. Methods Appl. Sci. 18, 1193–1215.
44. del Mar Delgado, M., Miranda, M., Alvarez, S.J., Gurarie, E., Fagan, W.F., Penteriani, V., di Virgilio, A., Morales, J.M., 2018. The importance of individual variation in the dynamics of animal collective movements. Philos. Trans. R. Soc. B Biol. Sci. 373, 20170008.
45. Di Francesco, M., Esposito, A., Fagioli, S., 2018. Nonlinear degenerate cross-diffusion systems with nonlocal interaction. Nonlinear Anal. 169, 94–117.
46. Di Francesco, M., Fagioli, S., 2013. Measure solutions for non-local interaction PDEs with two species. Nonlinearity 26, 2777.
47. Dorigo, M., Theraulaz, G., Trianni, V., 2020. Reflections on the future of swarm robotics. Sci. Robot. 5, eabe4385.
48. Dubois, D.M., 1975. A model of patchiness for prey—predator plankton populations. Ecol. Modell. 1, 67–80.
49. Dyer, J.R.G., Croft, D.P., Morrell, L.J., Krause, J., 2009. Shoal composition determines foraging success in the guppy. Behav. Ecol. 20, 165–171.
50. Edelstein-Keshet, L., 2001. Mathematical models of swarming and social aggregation, in: Proceedings of the 2001 International Symposium on Nonlinear Theory and Its Applications, Miyagi, Japan. Citeseer, pp. 1–7.
51. Faria, J.J., Krause, S., Krause, J., 2010. Collective behavior in road crossing pedestrians: the role of social information. Behav. Ecol. 21, 1236–1242.
52. Farine, D.R., Strandburg-Peshkin, A., Couzin, I.D., Berger-Wolf, T.Y., Crofoot, M.C., 2017. Individual variation in local interaction rules can explain emergent patterns of spatial organization in wild baboons. Proc. R. Soc. B Biol. Sci. 284, 20162243.
53. Feinerman, O., Pinkoviezky, I., Gelblum, A., Fonio, E., Gov, N.S., 2018. The physics of cooperative transport in groups of ants. Nat. Phys. 14, 683–693.

54. Flack, A., Nagy, M., Fiedler, W., Couzin, I.D., Wikelski, M., 2018. From local collective behavior to global migratory patterns in white storks. Science (80-.). 360, 911–914.
55. Fonio, E., Heyman, Y., Boczkowski, L., Gelblum, A., Kosowski, A., Korman, A., Feinerman, O., 2016. A locally-blazed ant trail achieves efficient collective navigation despite limited information. Elife 5, e20185.
56. Frouvelle, A., 2012. A continuum model for alignment of self-propelled particles with anisotropy and density-dependent parameters. Math. Model. Methods Appl. Sci. 22, 1250011.
57. Garland, J., Berdahl, A.M., Sun, J., Bollt, E.M., 2018. Anatomy of leadership in collective behaviour. Chaos An Interdiscip. J. Nonlinear Sci. 28, 75308.
58. Gavagnin, E., Yates, C.A., 2018. Stochastic and deterministic modeling of cell migration, in: Handbook of Statistics. Elsevier, pp. 37–91.
59. Gelblum, A., Pinkoviezky, I., Fonio, E., Ghosh, A., Gov, N., Feinerman, O., 2015. Ant groups optimally amplify the effect of transiently informed individuals. Nat. Commun. 6, 1–9.
60. Georgiou, F.H., Buhl, J., Green, J.E.F., Lamichhane, B., Thamwattana, N., 2020. Modelling locust foraging: How and why food affects hopper band formation. bioRxiv.
61. Giardina, I., 2008. Collective behavior in animal groups: theoretical models and empirical studies. HFSP J. 2, 205–219.
62. Gopalsamy, K., 1977. Competition and coexistence in spatially heterogeneous environments. Math. Biosci. 36, 229–242.
63. Gorbonos, D., Gov, N.S., 2017. Stable swarming using adaptive long-range interactions. Phys. Rev. E 95, 42405.
64. Gosling, S.D., 2001. From mice to men: what can we learn about personality from animal research? Psychol. Bull. 127, 45.
65. Grafke, T., Cates, M.E., Vanden-Eijnden, E., 2017. Spatiotemporal self-organization of fluctuating bacterial colonies. Phys. Rev. Lett. 119, 188003.
66. Guisandez, L., Baglietto, G., Rozenfeld, A., 2017. Heterogeneity promotes first to second order phase transition on flocking systems. arXiv Prepr. arXiv1711.11531.
67. Gurtin, M.E., Pipkin, A.C., 1984. A note on interacting populations that disperse to avoid crowding. Q. Appl. Math. 42, 87–94.
68. Ha, S.-Y., Tadmor, E., 2008. From particle to kinetic and hydrodynamic descriptions of flocking. Kinet. Relat. Model. 1, 415.
69. Helbing, D., 2001. Traffic and related self-driven many-particle systems. Rev. Mod. Phys. 73, 1067.
70. Hemelrijk, C.K., Hildenbrandt, H., 2011. Some causes of the variable shape of flocks of birds. PLoS One 6, e22479.
71. Hemelrijk, C.K., Hildenbrandt, H., 2008. Self-organized shape and frontal density of fish schools. Ethology 114, 245–254.
72. Hemelrijk, C.K., Kunz, H., 2005. Density distribution and size sorting in fish schools: an individual-based model. Behav. Ecol. 16, 178–187.
73. Herbert-Read, J.E., Krause, S., Morrell, L.J., Schaerf, T.M., Krause, J., Ward, A.J.W., 2013. The role of individuality in collective group movement. Proc. R. Soc. B Biol. Sci. 280, 20122564.
74. Herbert-Read, J.E., Rosén, E., Szorkovszky, A., Ioannou, C.C., Rogell, B., Perna, A., Ramnarine, I.W., Kotrschal, A., Kolm, N., Krause, J., 2017. How predation shapes the social interaction rules of shoaling fish. Proc. R. Soc. B Biol. Sci. 284, 20171126.
75. Hibbing, M.E., Fuqua, C., Parsek, M.R., Peterson, S.B., 2010. Bacterial competition: surviving and thriving in the microbial jungle. Nat. Rev. Microbiol. 8, 15–25.
76. Hinz, R.C., de Polavieja, G.G., 2017. Ontogeny of collective behavior reveals a simple attraction rule. Proc. Natl. Acad. Sci. 114, 2295–2300.
77. Horn, H.S., MacArthur, R.H., 1972. Competition among fugitive species in a harlequin environment. Ecology 53, 749–752.
78. Ihle, T., 2011. Kinetic theory of flocking: Derivation of hydrodynamic equations. Phys. Rev. E 83, 30901.

79. Ilkanaiv, B., Kearns, D.B., Ariel, G., Be'er, A., 2017. Effect of cell aspect ratio on swarming bacteria. Phys. Rev. Lett. 118, 158002.
80. Ingham, C.J., Kalisman, O., Finkelshtein, A., Ben-Jacob, E., 2011. Mutually facilitated dispersal between the nonmotile fungus Aspergillus fumigatus and the swarming bacterium Paenibacillus vortex. Proc. Natl. Acad. Sci. 108, 19731–19736.
81. Ioannou, C.C., Guttal, V., Couzin, I.D., 2012. Predatory fish select for coordinated collective motion in virtual prey. Science (80-.). 337, 1212–1215.
82. Jacoby, D.M.P., Papastamatiou, Y.P., Freeman, R., 2016. Inferring animal social networks and leadership: applications for passive monitoring arrays. J. R. Soc. Interface 13, 20160676.
83. Jeckel, H., Jelli, E., Hartmann, R., Singh, P.K., Mok, R., Totz, J.F., Vidakovic, L., Eckhardt, B., Dunkel, J., Drescher, K., 2019. Learning the space-time phase diagram of bacterial swarm expansion. Proc. Natl. Acad. Sci. 116, 1489–1494.
84. Jolles, J.W., Boogert, N.J., Sridhar, V.H., Couzin, I.D., Manica, A., 2017. Consistent individual differences drive collective behavior and group functioning of schooling fish. Curr. Biol. 27, 2862–2868.
85. Jolles, J.W., King, A.J., Killen, S.S., 2020. The role of individual heterogeneity in collective animal behaviour. Trends Ecol. Evol. 35, 278–291.
86. Jolles, J.W., Laskowski, K.L., Boogert, N.J., Manica, A., 2018. Repeatable group differences in the collective behaviour of stickleback shoals across ecological contexts. Proc. R. Soc. B Biol. Sci. 285, 20172629.
87. Kai, M., Piechulla, B., 2018. Interspecies interaction of Serratia plymuthica 4Rx13 and Bacillus subtilis B2g alters the emission of sodorifen. FEMS Microbiol. Lett.
88. Kareiva, P., 1990. Population dynamics in spatially complex environments: theory and data. Philos. Trans. R. Soc. London. Ser. B Biol. Sci. 330, 175–190.
89. Katz, Y., Tunstrøm, K., Ioannou, C.C., Huepe, C., Couzin, I.D., 2011. Inferring the structure and dynamics of interactions in schooling fish. Proc. Natl. Acad. Sci. 108, 18720–18725.
90. Kearns, D.B., Losick, R., 2005. Cell population heterogeneity during growth of Bacillus subtilis. Genes Dev. 19, 3083–3094.
91. Kennedy, J., Eberhart, R., 1995. Particle swarm optimization, in: Proceedings of ICNN'95-International Conference on Neural Networks. IEEE, pp. 1942–1948.
92. Kerr, B., Riley, M.A., Feldman, M.W., Bohannan, B.J.M., 2002. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. Nature 418, 171–174.
93. Khodygo, V., Swain, M.T., Mughal, A., 2019. Homogeneous and heterogeneous populations of active rods in two-dimensional channels. Phys. Rev. E 99, 22602.
94. Knebel, D., Ayali, A., Guershon, M., Ariel, G., 2019. Intra-versus intergroup variance in collective behavior. Sci. Adv. 5, eaav0695.
95. Knebel, D., Sha-Ked, C., Agmon, N., Ariel, G., Ayali, A., 2021. Collective motion as a distinct behavioral state of the individual. iScience 24, 102299.
96. Kotrschal, A., Szorkovszky, A., Herbert-Read, J., Bloch, N.I., Romenskyy, M., Buechel, S.D., Eslava, A.F., Alòs, L.S., Zeng, H., Le Foll, A., 2020. Rapid evolution of coordinated and collective movement in response to artificial selection. Sci. Adv. 6, eaba3148.
97. Krause, J., Ruxton, G.D., Ruxton, G., Ruxton, I.G., 2002. Living in groups. Oxford University Press.
98. Kumar, N., Soni, H., Ramaswamy, S., Sood, A.K., 2014. Flocking at a distance in active granular matter. Nat. Commun. 5, 1–9.
99. Kunz, H., Hemelrijk, C.K., 2003. Artificial fish schools: collective effects of school size, body size, and body form. Artif. Life 9, 237–253.
100. Leca, J.-B., Gunst, N., Thierry, B., Petit, O., 2003. Distributed leadership in semifree-ranging white-faced capuchin monkeys. Anim. Behav. 66, 1045–1052.
101. Levin, S.A., 1976. Population dynamic models in heterogeneous environments. Annu. Rev. Ecol. Syst. 287–310.
102. Lewis, J.S., Wartzok, D., Heithaus, M.R., 2011. Highly dynamic fission–fusion species can exhibit leadership when traveling. Behav. Ecol. Sociobiol. 65, 1061–1069.

103. Ling, H., Mclvor, G.E., van der Vaart, K., Vaughan, R.T., Thornton, A., Ouellette, N.T., 2019. Costs and benefits of social relationships in the collective motion of bird flocks. Nat. Ecol. Evol. 3, 943–948.
104. May, R.M., 1974. Ecosystem patterns in randomly fluctuating environments. Prog. Theor. Biol. 1–50.
105. McCandlish, S.R., Baskaran, A., Hagan, M.F., 2012. Spontaneous segregation of self-propelled particles with different motilities. Soft Matter 8, 2527–2534.
106. McComb, K., Shannon, G., Durant, S.M., Sayialel, K., Slotow, R., Poole, J., Moss, C., 2011. Leadership in elephants: the adaptive value of age. Proc. R. Soc. B Biol. Sci. 278, 3270–3276.
107. McMurtrie, R., 1978. Persistence and stability of single-species and prey-predator systems in spatially heterogeneous environments. Math. Biosci. 39, 11–51.
108. Menzel, A.M., 2012. Collective motion of binary self-propelled particle mixtures. Phys. Rev. E 85, 21912.
109. Mimura, M., Kawasaki, K., 1980. Spatial segregation in competitive interaction-diffusion equations. J. Math. Biol. 9, 49–64.
110. Mishra, S., Tunstrøm, K., Couzin, I.D., Huepe, C., 2012. Collective dynamics of self-propelled particles with variable speed. Phys. Rev. E 86, 11901.
111. Munson, A., Michelangeli, M., Sih, A., 2021. Stable social groups foster conformity and among-group differences. Anim. Behav. 174, 197–206.
112. Nadell, C.D., Drescher, K., Foster, K.R., 2016. Spatial structure, cooperation and competition in biofilms. Nat. Rev. Microbiol. 14, 589–600.
113. Nagy, M., Ákos, Z., Biro, D., Vicsek, T., 2010. Hierarchical group dynamics in pigeon flocks. Nature 464, 890–893.
114. Namba, T., 1989. Competition for space in a heterogeneous environment. J. Math. Biol. 27, 1–16.
115. Namba, T., 1980. Density-dependent dispersal and spatial distribution of a population. J. Theor. Biol. 86, 351–363.
116. Namba, T., Mimura, M., 1980. Spatial distribution of competing populations. J. Theor. Biol. 87, 795–814.
117. Navoret, L., 2013. A two-species hydrodynamic model of particles interacting through self-alignment. Math. Model. Methods Appl. Sci. 23, 1067–1098.
118. Netzer, G., Yarom, Y., Ariel, G., 2019. Heterogeneous populations in a network model of collective motion. Phys. A Stat. Mech. its Appl. 530, 121550.
119. Peled, S., Ryan, S.D., Heidenreich, S., Bär, M., Ariel, G., Be'er, A., 2021. Heterogeneous bacterial swarms with mixed lengths. Phys. Rev. E 103, 32413.
120. Penaz, M., 2001. A general framework of fish ontogeny: a review of the ongoing debate. Folia Zool. (Czech Republic).
121. Pettit, B., Akos, Z., Vicsek, T., Biro, D., 2015. Speed determines leadership and leadership determines learning during pigeon flocking. Curr. Biol. 25, 3132–3137.
122. Planas-Sitjà, I., Deneubourg, J.-L., Cronin, A.L., 2021. Variation in personality can substitute for social feedback in coordinated animal movements. Commun. Biol. 4, 1–13.
123. Porfiri, M., Ariel, G., 2016. On effective temperature in network models of collective behavior. Chaos An Interdiscip. J. Nonlinear Sci. 26, 43109.
124. Rahmani, P., Peruani, F., Romanczuk, P., 2021. Topological flocking models in spatially heterogeneous environments. Commun. Phys. 4, 1–9.
125. Reches, E., Knebel, D., Rillich, J., Ayali, A., Barzel, B., 2019. The metastability of the double-tripod gait in locust locomotion. Iscience 12, 53–65.
126. Reichenbach, T., Mobilia, M., Frey, E., 2007. Mobility promotes and jeopardizes biodiversity in rock–paper–scissors games. Nature 448, 1046–1049.
127. Reynolds, C.W., 1987. Flocks, herds and schools: A distributed behavioral model, in: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. pp. 25–34.
128. Roff, D.A., 1974a. Spatial heterogeneity and the persistence of populations. Oecologia 15, 245–258.

129. Roff, D.A., 1974b. The analysis of a population model demonstrating the importance of dispersal in a heterogeneous environment. Oecologia 15, 259–275.
130. Romey, W.L., 1997. Inside or outside? Testing evolutionary predictions of positional effects. Anim. groups three Dimens. 174–193.
131. Rørth, P., 2012. Fellow travellers: emergent properties of collective cell migration. EMBO Rep. 13, 984–991.
132. Rosenberg, G., Steinberg, N., Oppenheimer-Shaanan, Y., Olender, T., Doron, S., Ben-Ari, J., Sirota-Madi, A., Bloom-Ackermann, Z., Kolodkin-Gal, I., 2016. Not so simple, not so subtle: the interspecies competition between Bacillus simplex and Bacillus subtilis and its impact on the evolution of biofilms. NPJ biofilms microbiomes 2, 1–11.
133. Roughgarden, J., 1974. Population dynamics in a spatially varying environment: how population size "tracks" spatial variation in carrying capacity. Am. Nat. 108, 649–664.
134. Ryan, S.D., Haines, B.M., Berlyand, L., Ziebert, F., Aranson, I.S., 2011. Viscosity of bacterial suspensions: Hydrodynamic interactions and self-induced noise. Phys. Rev. E 83, 50904.
135. Schumacher, L.J., Kulesa, P.M., McLennan, R., Baker, R.E., Maini, P.K., 2016. Multidisciplinary approaches to understanding collective cell migration in developmental biology. Open Biol. 6, 160056.
136. Schumacher, L.J., Maini, P.K., Baker, R.E., 2017. Semblance of heterogeneity in collective cell migration. Cell Syst. 5, 119–127.
137. Schweitzer, F., Schimansky-Geier, L., 1994. Clustering of "active" walkers in a two-component system. Phys. A Stat. Mech. its Appl. 206, 359–379.
138. Segel, L.A., Levin, S.A., 1976. Application of nonlinear stability theory to the study of the effects of diffusion on predator-prey interactions, in: AIP Conference Proceedings. American Institute of Physics, pp. 123–152.
139. Shaukat, M., Chitre, M., 2016. Adaptive behaviors in multi-agent source localization using passive sensing. Adapt. Behav. 24, 446–463.
140. Shigesada, N., Kawasaki, K., Teramoto, E., 1979. Spatial segregation of interacting species. J. Theor. Biol. 79, 83–99.
141. Shklarsh, A., Ariel, G., Schneidman, E., Ben-Jacob, E., 2011. Smart swarms of bacteria-inspired agents with performance adaptable interactions. PLoS Comput Biol 7, e1002177.
142. Sih, A., 1980. Optimal behavior: can foragers balance two conflicting demands? Science (80-.). 210, 1041–1043.
143. Singh, J.P., Mishra, S., 2020. Phase separation in a binary mixture of self-propelled particles with variable speed. Phys. A Stat. Mech. its Appl. 544, 123530.
144. Smith, J.E., Gavrilets, S., Mulder, M.B., Hooper, P.L., El Mouden, C., Nettle, D., Hauert, C., Hill, K., Perry, S., Pusey, A.E., 2016. Leadership in mammalian societies: Emergence, distribution, power, and payoff. Trends Ecol. Evol. 31, 54–66.
145. Soni, H., Kumar, N., Nambisan, J., Gupta, R.K., Sood, A.K., Ramaswamy, S., 2020. Phases and excitations of active rod–bead mixtures: simulations and experiments. Soft Matter 16, 7210–7221.
146. Stefanic, P., Kraigher, B., Lyons, N.A., Kolter, R., Mandic-Mulec, I., 2015. Kin discrimination between sympatric Bacillus subtilis isolates. Proc. Natl. Acad. Sci. 112, 14042–14047.
147. Strandburg-Peshkin, A., Farine, D.R., Couzin, I.D., Crofoot, M.C., 2015. Shared decision-making drives collective movement in wild baboons. Science (80-.). 348, 1358–1361.
148. Strandburg-Peshkin, A., Farine, D.R., Crofoot, M.C., Couzin, I.D., 2017. Habitat and social factors shape individual decisions and emergent group structure during baboon collective movement. Elife 6, e19505.
149. Sumpter, D.J.T., 2010. Collective animal behavior. Princeton University Press.
150. Szabo, B., Szöllösi, G.J., Gönci, B., Jurányi, Z., Selmeczi, D., Vicsek, T., 2006. Phase transition in the collective migration of tissue cells: experiment and model. Phys. Rev. E 74, 61908.
151. Tadmor, E., 2021. On the Mathematics of Swarming: Emergent Behavior in Alignment Dynamics. Not. AMS 68, 493–503.

152. Theodorakis, C.W., 1989. Size segregation and the effects of oddity on predation risk in minnow schools. Anim. Behav. 38, 496–502.
153. Tipping, M.J., Gibbs, K.A., 2019. Peer pressure from a Proteus mirabilis self-recognition system controls participation in cooperative swarm motility. PLoS Pathog. 15, e1007885.
154. Toner, J., Tu, Y., Ramaswamy, S., 2005. Hydrodynamics and phases of flocks. Ann. Phys. (N. Y). 318, 170–244.
155. Tong, H., Chen, W., Merritt, J., Qi, F., Shi, W., Dong, X., 2007. Streptococcus oligofermentans inhibits Streptococcus mutans through conversion of lactic acid into inhibitory H2O2: a possible counteroffensive strategy for interspecies competition. Mol. Microbiol. 63, 872–880.
156. Topaz, C.M., D'Orsogna, M.R., Edelstein-Keshet, L., Bernoff, A.J., 2012. Locust dynamics: behavioral phase change and swarming. PLoS Comput Biol 8, e1002642.
157. Torney, C., Neufeld, Z., Couzin, I.D., 2009. Context-dependent interaction leads to emergent search behavior in social aggregates. Proc. Natl. Acad. Sci. 106, 22055–22060.
158. Ulrich, Y., Saragosti, J., Tokita, C.K., Tarnita, C.E., Kronauer, D.J.C., 2018. Fitness benefits and emergent division of labour at the onset of group living. Nature 560, 635–638.
159. Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I., Shochet, O., 1995. Novel type of phase transition in a system of self-driven particles. Phys. Rev. Lett. 75, 1226.
160. Vicsek, T., Zafeiris, A., 2012. Collective motion. Phys. Rep. 517, 71–140.
161. Viscido, S. V, Parrish, J.K., Grünbaum, D., 2004. Individual behavior and emergent properties of fish schools: a comparison of observation and theory. Mar. Ecol. Prog. Ser. 273, 239–249.
162. Ward, A., Webster, M., 2016. Sociality: the behaviour of group-living animals.
163. Ward, A.J.W., Schaerf, T.M., Burns, A.L.J., Lizier, J.T., Crosato, E., Prokopenko, M., Webster, M.M., 2018. Cohesion, order and information flow in the collective motion of mixed-species shoals. R. Soc. open Sci. 5, 181132.
164. Watts, I., Nagy, M., Holbrook, R.I., Biro, D., Burt de Perera, T., 2017. Validating two-dimensional leadership models on three-dimensionally structured fish schools. R. Soc. open Sci. 4, 160804.
165. Wensink, H.H., Dunkel, J., Heidenreich, S., Drescher, K., Goldstein, R.E., Löwen, H., Yeomans, J.M., 2012. Meso-scale turbulence in living fluids. Proc. Natl. Acad. Sci. 109, 14308–14313.
166. Witelski, T.P., 1997. Segregation and mixing in degenerate diffusion in population dynamics. J. Math. Biol. 35, 695–712.
167. Wolf, M., Weissing, F.J., 2012. Animal personalities: consequences for ecology and evolution. Trends Ecol. Evol. 27, 452–461.
168. Xue, T., Li, X., Grassberger, P., Chen, L., 2020. Swarming transitions in hierarchical societies. Phys. Rev. Res. 2, 42017.
169. Yllanes, D., Leoni, M., Marchetti, M.C., 2017. How many dissenters does it take to disorder a flock? New J. Phys. 19, 103026.
170. Zuo, W., Wu, Y., 2020. Dynamic motility selection drives population segregation in a bacterial swarm. Proc. Natl. Acad. Sci. 117, 4693–4700.

Active Crowds

Maria Bruna, Martin Burger, Jan-Frederik Pietschmann, and Marie-Therese Wolfram

Abstract This chapter focuses on the mathematical modelling of active particles (or agents) in crowded environments. We discuss several microscopic models found in the literature and the derivation of the respective macroscopic partial differential equations for the particle density. The macroscopic models share common features, such as cross-diffusion or degenerate mobilities. We then take the diversity of macroscopic models to a uniform structure and work out potential similarities and differences. Moreover, we discuss boundary effects and possible applications in life and social sciences. This is complemented by numerical simulations that highlight the effects of different boundary conditions.

1 Introduction

The mathematical modelling of active matter has received growing interest recently, motivated by novel structures in physics and biology on the one hand (cf. [7, 34, 38, 40, 60, 66, 69]), but also active matter in a wider sense of agent systems like humans

M. Bruna
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
e-mail: bruna@maths.cam.ac.uk

M. Burger
Department Mathematik, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany
e-mail: martin.burger@fau.edu

J.-F. Pietschmann
Fakultät für Mathematik, Technische Universität, Chemnitz, Germany
e-mail: jfpietschmann@math.tu-chemnitz.de

M.-T. Wolfram (✉)
Mathematics Institute, University of Warwick, Coventry, UK

Radon Institute for Computational and Applied Mathematics, Linz, Austria
e-mail: m.wolfram@warwick.ac.uk

N. Bellomo et al. (eds.), *Active Particles, Volume 3*, Modeling and Simulation in Science, Engineering and Technology, https://doi.org/10.1007/978-3-030-93302-9_2

or robots (cf. [8, 22, 29, 30, 54, 53, 37]). In many of these systems, a key issue is the interplay of the particles' own activity with crowding effects, which leads to the formation of complex and interesting patterns. In this chapter, we aim at unifying a variety of models proposed for such systems, discuss the derivation of macroscopic equations from different microscopic paradigms and highlight some of their main properties.

First, we would like to emphasise that the definition of active particles varies in the literature, e.g. between condensed matter, life sciences and engineering and that we will adopt a generous point of view in this chapter. We will discuss not only models for actual active matter systems but also systems that might be considered passive (or force-driven) in the physics literature but have been used to model active matter systems. A common feature of these models is that the particles have a preferred direction of motion and can use energy to move there; the preferred direction can however change in time. This setting includes models for multiple species (with fixed directions or biases) and systems with boundary conditions that impose steady currents. From a mathematical point of view, we can distinguish whether models exhibit a gradient-flow structure (cf. [4]) or not, respectively, and whether there are stationary solutions with vanishing flux. In the case of a gradient-flow structure, there is a natural entropy–energy functional to be dissipated (cf. [25]), and in the other cases, such functionals may increase (linearly) in time or it is not apparent what the correct choice of the functional is. We shall see in particular that gradient-flow structures are destroyed if particles change directions completely randomly, while there is an active transport in that direction.

In the following, we will consider models with a finite number of preferred directions or a continuum of it. While the former has been proposed and investigated in many applications like pedestrian dynamics or cell motility, continuum directions received far less attention in the mathematical literature. In the continuum case, one can consider angular diffusion and derive an equation for the density of particles in the phase-space of spatial and angular variables. We discuss different microscopic models, based either on Brownian motions with hard sphere potentials or on lattice-based models with size exclusion, which allow to derive appropriate macroscopic partial differential equations (PDEs) for the phase-space density. These PDEs have a rather similar structure—all have a nonlinear transport term and additional diffusion terms in space and angle (or possibly nonlocal diffusion for the latter). This general structure allows us to define a general entropy functional and investigate the long-time behaviour.

This chapter is organised as follows: we present several microscopic models for active particles and their corresponding mean-field limits using different coarse-graining procedures in Sect. 2. Section 3 discusses the respective modelling approaches and limiting equations for externally activated particles (systems that would be considered passive in the physics literature). We then present a general formulation of all these models and state their underlying properties, such as energy dissipation or a possible underlying gradient-flow structure in Sect. 4. The important role of boundary conditions on the behaviour of these systems is discussed in Sect. 5. Section 6 presents several examples of active and externally activated particle models in the life and social sciences. Numerical experiments illustrating the dynamics and behaviour of the respective models are presented in Sect. 7.

Throughout this chapter, we use the notion of particles or agents interchangeably. Furthermore, we discuss the respective models on the line or in $\mathbb{R}^2$, with the obvious generalisation to three dimensions. We will keep the presentation informal, assuming that all functions satisfy the necessary requirements to perform all limits and calculations.

2 Models for Active Particles

Here, we discuss microscopic models for active particles and their corresponding macroscopic kinetic models using different coarse-graining procedures. The key ingredients of active particles are that, in addition to their positions, they have an orientation that determines the self-propulsion direction. We subdivide these models into continuous, discrete or hybrid random walks depending on how the position and the orientation of each particle are represented.

2.1 Continuous Random Walks

We consider N identical Brownian particles with free translational diffusion coefficient D_T moving in a periodic box $\Omega \subset \mathbb{R}^2$ with unit area. Each particle has a position $\mathbf{X}_i(t)$ and an orientation $\Theta_i(t)$ with $t > 0$, $i = 1, \ldots, N$, that determines the direction $\mathbf{e}(\Theta_i) = (\cos\Theta_i, \sin\Theta_i)$ of self-propulsion at constant speed v_0. The orientation Θ_i also undergoes free rotational diffusion with diffusion coefficient D_R. In its more general form, particles interact through a pair potential $u(r, \varphi)$ which implies the total potential energy

$$U = \chi \sum_{1 \le i < j \le N} u(|\mathbf{X}_i - \mathbf{X}_j|/\ell, |\Theta_i - \Theta_j|), \tag{1}$$

where χ and ℓ represent the strength and the range in space of the potential u, respectively. The coupled equations of motion are

$$\mathrm{d}\mathbf{X}_i = \sqrt{2D_T}\mathrm{d}\mathbf{W}_i - \nabla_{\mathbf{x}_i} U \mathrm{d}t + v_0 \mathbf{e}(\Theta_i)\mathrm{d}t, \tag{2a}$$

$$\mathrm{d}\Theta_i = \sqrt{2D_R}\mathrm{d}W_i - \partial_{\theta_i} U \mathrm{d}t. \tag{2b}$$

We note that isotropic pair potentials ($u = u(r)$ only) are more commonly used in the literature [10, 46]. System (2) is complemented with identically and independently distributed initial conditions, $(\mathbf{X}_i(0), \Theta_i(0)) \sim f_0(\mathbf{x}, \theta)$ and periodic boundary conditions on $\Upsilon = \Omega \times [0, 2\pi)$ (we will discuss alternative boundary conditions later in Sect. 5).

The starting point for all is to define the joint probability density for N particles evolving according to the SDEs (2), that is, $F_N(\vec{\xi}, t)$ with $\vec{\xi} = (\xi_1, \dots, \xi_N)$ and $\xi_i = (\mathbf{x}_i, \theta_i)$. Using the Chapman–Kolmogorov equation, we obtain

$$\begin{aligned}\partial_t F_N(\vec{\xi}, t) = \sum_{i=1}^{N} \nabla_{\mathbf{x}_i} \cdot \Big[D_T \nabla_{\mathbf{x}_i} F_N - v_0 \mathbf{e}(\theta_i) F_N + \nabla_{\mathbf{x}_i} U(\vec{\xi}) F_N \Big] \\ + \partial_{\theta_i} \Big[D_R \partial_{\theta_i} F_N + \partial_{\theta_i} U(\vec{\xi}) F_N \Big], \end{aligned} \tag{3}$$

for $t \geq 0, \vec{\xi} \in \bar{\Upsilon}$, where $\bar{\Upsilon} \subseteq \Upsilon^N$ is the domain of allowed configurations (more on this below).

The goal is to obtain a macroscopic model for the one-particle density $f(\xi, t)$, which we can describe by picking the first particle since all particles are identical, i.e.

$$f(\xi_1, t) = \int_{\Upsilon^N} F_N(\vec{\xi}) \delta(\xi - \xi_1) \mathrm{d}\vec{\xi}. \tag{4}$$

To this end, keeping in mind all the particles are indistinguishable, we integrate (3) with respect to $\xi_2, \dots, \xi_N$. Using periodicity, all the terms for $i \geq 2$ in the right-hand side of (3) vanish, and we are left with

$$\partial_t f(\xi_1, t) = \nabla_{\mathbf{x}_1} \cdot \big[D_T \nabla_{\mathbf{x}_1} f - v_0 \mathbf{e}(\theta_1) f + \mathbf{U}_T(\xi_1, t) \big] + \partial_{\theta_1} \big[D_R \partial_{\theta_1} f + U_R(\xi_1, t) \big], \tag{5a}$$

with

$$\begin{aligned}\mathbf{U}_T(\xi_1, t) &= \chi \int_{\Upsilon^{N-1}} F_N(\vec{\xi}, t) \sum_{i=2}^{N} \nabla_{\mathbf{x}_1} u(|\mathbf{x}_1 - \mathbf{x}_i|/\ell, |\theta_1 - \theta_i|) \mathrm{d}\xi_2, \dots, \mathrm{d}\xi_N \\ &= \chi(N-1) \int_{\Upsilon} F_2(\xi_1, \xi_2, t) \nabla_{\mathbf{x}_1} u(|\mathbf{x}_1 - \mathbf{x}_2|/\ell, |\theta_1 - \theta_2|) \mathrm{d}\xi_2, \end{aligned} \tag{5b}$$

$$\begin{aligned}U_R(\xi_1, t) &= \chi \int_{\Upsilon^{N-1}} F_N(\vec{\xi}, t) \sum_{i=2}^{N} \partial_{\theta_1} u(|\mathbf{x}_1 - \mathbf{x}_i|/\ell, |\theta_1 - \theta_i|) \mathrm{d}\xi_2, \dots, \mathrm{d}\xi_N \\ &= \chi(N-1) \int_{\Upsilon} F_2(\xi_1, \xi_2, t) \partial_{\theta_1} u(|\mathbf{x}_1 - \mathbf{x}_2|/\ell, |\theta_1 - \theta_2|) \mathrm{d}\xi_2, \end{aligned} \tag{5c}$$

using that particles are undistinguishable, where F_2 is the two-particle density

$$F_2(\xi_1, \xi_2, t) = \int_{\Upsilon^{N-2}} F_N(\vec{\xi}, t)\mathrm{d}\xi_3 \dots \mathrm{d}\xi_N.$$

Depending on the scalings χ, ℓ of the interaction potential u, we can expect different macroscopic limits of (2). On the one end, we can consider long-range weak repulsive interactions and obtain a mean-field limit equation. On the other extreme, we can consider short and strong repulsive interactions (even hard-core interactions such as $u(r) = +\infty$ if $r < 1$, and 0 otherwise), which lead to local nonlinear PDE models.

Mean-Field Scaling

The mean-field scaling corresponds to $\chi = 1/N$ and $\ell = O(1)$ so that we have a weak and long-range interaction in the limit of $N \to \infty$. In this limit, one expects propagation of chaos leading to

$$F_2(\xi_1, \xi_2, t) \approx f(\xi_1, t) f(\xi_2, t).$$

Substituting this into (5), we arrive at

$$\partial_t f(\xi_1, t) = \nabla_{\mathbf{x}_1} \cdot \left[D_T \nabla_{\mathbf{x}_1} f - v_0 \mathbf{e}(\theta_1) f + f \nabla_{\mathbf{x}_1} \mathcal{U}\right] + \partial_{\theta_1} \left[D_R \partial_{\theta_1} f + \partial_{\theta_1} \mathcal{U}\right], \tag{6a}$$

with interaction term, taking $N \to \infty$,

$$\mathcal{U}(f) = \int_{\Upsilon} f(\xi_2, t) u(|\mathbf{x}_1 - \mathbf{x}_2|/\ell, |\theta_1 - \theta_2|)\mathrm{d}\xi_2. \tag{6b}$$

In the case of an isotropic interaction potential, the term $\partial_{\theta_1}\mathcal{U}$ in (6a) drops, and the interaction term (6b) can be simplified to

$$\mathcal{U}(f) = \int_{\Omega} \rho(\mathbf{x}_2, t) u(|\mathbf{x}_1 - \mathbf{x}_2|/\ell)\mathrm{d}\mathbf{x}_2, \tag{7}$$

where ρ is the spatial (macroscopic) density

$$\rho(\mathbf{x}, t) = \int_0^{2\pi} f(\mathbf{x}, \theta, t)\, \mathrm{d}\theta. \tag{8}$$

The spatial density describes the probability that a particle is at position $\mathbf{x}$ at time t irrespective of its orientation. We obtain the following equation for ρ by integrating (6a) with the potential (7) with respect to $\theta \in [0, 2\pi)$ and using periodicity:

$$\partial_t \rho(\mathbf{x}_1, t) = \nabla_{\mathbf{x}_1} \cdot \left[D_T \nabla_{\mathbf{x}_1} \rho - v_0 \mathbf{p} + \rho \nabla_{\mathbf{x}_1} \mathcal{U}\right]. \tag{9}$$

This equation is not closed as it depends on the polarisation $\mathbf{p}$ (also known as the order parameter):

$$\mathbf{p}(\mathbf{x}, t) = \int_0^{2\pi} \mathbf{e}(\theta) f(\mathbf{x}, \theta, t)\, \mathrm{d}\theta. \tag{10}$$

The polarisation gives the average orientation of particles at position $\mathbf{x}$ at any given time t.

2.1.1 Excluded-Volume Interactions

Excluded-volume interactions are very common in biological applications and arise from the impenetrability between cells, bacteria, animals etc. These are very strong and short-range interactions, whereby an individual only interacts locally in the range of its body size. For these reasons, the mean-field scaling is not suitable to model such interactions, which are often modelled using singular short-range potentials ($\ell \ll 1$ in (1)) or even hard-core potentials. Examples of interaction potentials used in the literature to model excluded-volume interactions include inverse power-law potentials (such as the Lennard-Jones potentials), exponential potentials (e.g. the Morse potential) or the Yukawa potential.

The following model, proposed by [65], includes excluded-volume interactions via a short-range interaction potential $u(r)$:

$$\partial_t f + \nabla \cdot (v_e(\rho) f \mathbf{e}(\theta)) = D_e(\phi)\Delta f + D_R \partial_\theta^2 f, \tag{11}$$

where $f = f(r, \theta, t)$, $D_e(\phi)$ is an effective diffusion depending on how crowded the system is (given by ϕ), and $v_e = v_0(1 - \phi\rho)$ is a nonlinear effective speed. The hydrodynamic equations for the spatial density ρ and the polarisation $\mathbf{p}$ are obtained by integrating (11),

$$\partial_t \rho + \nabla \cdot (v_e(\rho)\mathbf{p}) = D_e(\phi)\Delta\rho, \tag{12}$$

$$\partial_t \mathbf{p} + \nabla P(\rho) = D_e(\phi)\Delta\mathbf{p} - \mathbf{p}, \tag{13}$$

with the so-called pressure $P(\rho) = v_e(\rho)\rho/2$. This model displays a motility-induced phase transition [65]: at low densities ($\phi\rho$ small), the effective swimming speed is close to the free speed v_0, whereas at high densities, the effective swimming goes to zero. The result is a phase separation, with regions of high density where particles are trapped and do not move, and very dilute areas with fast speeds. This is shown via a linear stability analysis as well as numerical simulations of the microscopic system using the repulsive Weeks–Chandler–Andersen (WCA) potential (which corresponds to a truncated and shifted upwards Lennard-Jones potential). Through an adiabatic approximation, they cast equation (12) into a gradient flow of an effective free energy of the form of a conventional Ginzburg–

Landau function. According to [65], this is consistent with 'the mapping of active phase separation onto that of passive fluids with attractive interactions through a global effective free energy'.

An alternative derivation of a macroscopic model for active Brownian particles is considered in [17] using the hard-core interaction potential, $u(r) = +\infty$ for $r < \epsilon$ and 0 otherwise. In this case, the microscopic model changes from (2) to

$$\mathrm{d}\mathbf{X}_i = \sqrt{2D_T}\mathrm{d}\mathbf{W}_i + v_0\mathbf{e}(\Theta_i)\mathrm{d}t, \qquad |\mathbf{X}_i - \mathbf{X}_i| > \epsilon, \forall j \neq i, \tag{14a}$$

$$\mathrm{d}\Theta_i = \sqrt{2D_R}\mathrm{d}W_i. \tag{14b}$$

This represents particles as hard disks of diameter ϵ: particles only sense each other when they come into contact, and they are not allowed to get closer than ϵ to each other (mutual impenetrability condition). In comparison with the mean-field scaling, here instead the scaling is $\chi = 1, \ell = \epsilon \ll 1$ so that each particle only interacts with the few particles that are within a distance $O(\epsilon)$, the interaction is very strong. Using the method of matched asymptotics, from (14), one obtains to order ϕ the following model:

$$\partial_t f + v_0 \nabla \cdot [f(1 - \phi\rho)\mathbf{e}(\theta) + \phi\mathbf{p}f] = D_T \nabla \cdot [(1 - \phi\rho)\nabla f + 3\phi f \nabla\rho] + D_R \partial_\theta^2 f. \tag{15}$$

Here, ϕ is the effective occupied area $\phi = (N-1)\epsilon^2\pi/2$. Model (15) is obtained formally in the limit of ϵ and ϕ small. Note that this equation is consistent with the case $N = 1$: if there is only one particle, then $\phi = 0$ and we recover a linear PDE (no interactions). The equation for the spatial density is

$$\partial_t \rho + v_0 \nabla \cdot \mathbf{p} = D_T \nabla \cdot [(1 + 2\phi\rho)\nabla\rho], \tag{16}$$

which indicates the collective diffusion effect: the higher the occupied fraction ϕ, the higher the effective diffusion coefficient. We note that, due to the nature of the excluded-volume interactions, models (11) and (15) are obtained via approximations (closure at the pair correlation function and matched asymptotic expansions, respectively) and no rigorous results are available. A nice exposition of the difficulties of going from micro to macro in the presence of hard-core non-overlapping constraints is given in [52]. In particular, they consider hard-core interacting particles in the context of congestion handling in crowd motion. In contrast to (14), the dynamics involve only position and are deterministic. Collisions can then be handled via the projection of velocities onto the set of feasible velocities. In [52], they do not attempt to derive a macroscopic model from the microscopic dynamics but instead propose a PDE model for the population density $\rho(\mathbf{x}, t)$ that expresses the congestion assumption by setting the velocity to zero whenever ρ attains a saturation value (which they set to one).

2.2 Discrete Random Walks

Next, we discuss fully discrete models for active particles with size exclusion effects. We start by considering a simple exclusion model for active particles on a one-dimensional lattice, which has been investigated in [48]. The brief description of the microscopic lattice model is as follows: N particles of size ϵ evolve on a discrete ring of $1/\epsilon$ sites, with occupancy $\phi = \epsilon N \leq 1$. Each lattice is occupied by at most one particle (thus modelling a size exclusion), and particles can either be moving left ($-$ particles) or right ($+$ particles). The configuration can be represented using occupation numbers σ_i at site i with values in $\{-1, 0, 1\}$. The dynamics combine three mechanisms:

(a) Diffusive motion: for each bond $(i, i+1)$, σ_i and σ_{i+1} are exchanged at rate $D_T \backslash \epsilon^2$.

(b) Self-propulsion and size exclusion: for each bond $(i, i+1)$, a $+$ particle in i jumps to $i+1$ if $\sigma_{i+1} = 0$, or a $-$ particle in $i+1$ jumps to i if $\sigma_i = 0$, both at rate ϵv_0.

(c) Tumbling: particles switch direction $\sigma_i \to -\sigma_i$ at rate $\epsilon^2 \lambda$;

see Fig. 1 for an illustration of these effects. Rescaling space and time as ϵi and $\epsilon^2 t$, respectively, and a smooth initial condition, the macroscopic equations can be derived exactly as [48]

$$
\begin{aligned}
\partial_t f_+ + v_0 \partial_x [f_+(1 - \phi\rho)] &= D_T \partial_{xx} f_+ + \lambda(f_- - f_+), \\
\partial_t f_- - v_0 \partial_x [f_-(1 - \phi\rho)] &= D_T \partial_{xx} f_- + \lambda(f_+ - f_-),
\end{aligned}
\tag{17}
$$

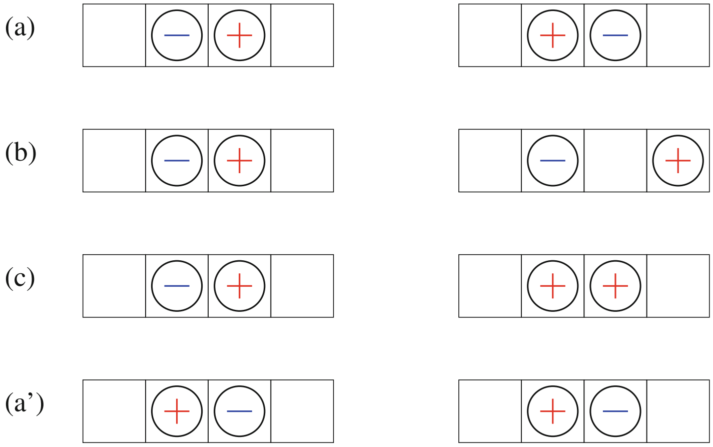


Fig. 1 Sketch illustrating the update steps for $+$ (right moving) and $-$ (left moving) particles outlined in Sect. 2.2. The left column shows the initial set-up and the right one the configuration after a single time step

where f_+ and f_- are the probability densities corresponding to the $+$ and $-$ particles, respectively, and $\rho = f_+ + f_-$. Introducing the number densities

$$r(\mathbf{x}, t) = N f_+(\mathbf{x}, t), \qquad b(\mathbf{x}, t) = N f_-(\mathbf{x}, t), \tag{18}$$

which integrate to N_1 and N_2, respectively, we can rewrite (17) as

$$\begin{aligned} \partial_t r + v_0 \partial_x [r(1 - \bar{\rho})] &= D_T \partial_{xx} r + \lambda(b - r), \\ \partial_t b - v_0 \partial_x [b(1 - \bar{\rho})] &= D_T \partial_{xx} b + \lambda(r - b), \end{aligned} \tag{19}$$

with $\bar{\rho} = \epsilon(r+b)$. One can also consider the same process in higher dimensions with a finite set of orientations $\mathbf{e}_k$, $k = 1, \ldots, m$. The most straightforward generalisation of (17) is to consider a two-dimensional square lattice with $m = 4$ directions, namely $\mathbf{e}_1 = (1, 0)$, $\mathbf{e}_2 = (0, 1)$, $\mathbf{e}_3 = (-1, 0)$ and $\mathbf{e}_4 = (0, -1)$ (see Fig. 1 in [48]). In this case, the configuration would take five possible values, $\sigma_i = \{-1, -i, 0, i, 1\}$, and the resulting macroscopic model would consist of a system of four equations for the densities of each subpopulations

$$\partial_t f_k + v_0 \nabla \cdot [f_k(1 - \phi\rho)\mathbf{e}_k] = D_T \Delta f_k + \lambda(f_{k+1} + f_{k-1} - 2 f_k), \qquad k = 1, \ldots, 4, \tag{20}$$

where now $\phi = \epsilon^2 N$, $f_k(\mathbf{x}, t)$ stands for the probability density of particles going in the $\mathbf{e}_k$ direction, and $\rho = \sum_k f_k$. Periodicity in angle implies that $f_5 = f_1$, $f_{-1} = f_4$.

Note how the model in [48] differs from an asymmetric simple exclusion processes (ASEP) in that particles are allowed to swap places in the diffusive step (see (a) above). As a result, the macroscopic models (17) and (20) lack any cross-diffusion terms. We can also consider an actual ASEP process, in which simple exclusion is also added to the diffusive step, that is, point (a) above is replaced by

(a′) Diffusive motion: a particle in i jumps to $i + 1$ at rate $D_T \backslash \epsilon^2$ if $\sigma_{i+1} = 0$ (and similarly to $i - 1$).

In this case, the resulting macroscopic model is

$$\begin{aligned} \partial_t f_k + v_0 \nabla \cdot [f_k(1 - \phi\rho)\mathbf{e}_k] &= D_T \nabla \cdot [(1 - \phi\rho)\nabla f_k + \phi f_k \nabla \rho] \\ &\quad + \lambda(f_{k+1} + f_{k-1} - 2 f_k), \quad k = 1, \ldots, 4. \end{aligned} \tag{21}$$

2.3 Hybrid Random Walks

In the previous two subsections, we have discussed models that consider both the position and the orientation as continuous or discrete. Here, we discuss hybrid random walks, that is, when positions are continuous and orientations finite, or vice versa.

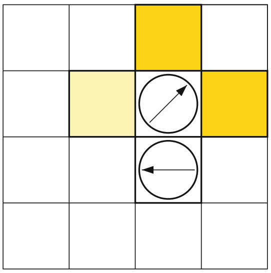


Fig. 2 Sketch of the 2D hybrid random walk outlined in Sect. 2.3. The arrow within each particle corresponds to its orientation, and the colour of the neighbouring sites relates to π_k. The darker the colour, the greater the likelihood to jump into the cell

The first hybrid model we consider is an active exclusion process whereby the orientation is a continuous process in $[0, 2\pi)$ evolving according to a Brownian motion with diffusion D_R, (2b), while keeping the position evolving according to a discrete asymmetric exclusion process (ASEP) [17]. The advantage of this approach is to avoid the anisotropy imposed by the underlying lattice. Here, we present the model in two dimensions so that we can compare it to the models presented above.

We consider a square lattice with spacing ϵ and orientations $\mathbf{e}_k, k = 1, \ldots, 4$ as given above. A particle at lattice site $\mathbf{x}$ can jump to neighbouring sites $\mathbf{x} + \epsilon\mathbf{e}_k$ if the latter is empty at a rate $\pi_k(\theta)$ that depends on its orientation θ, namely

$$\pi_k(\theta) = \alpha_\epsilon \exp(\beta_\epsilon \mathbf{e}(\theta) \cdot \mathbf{e}_k),$$

where $\alpha_\epsilon = D_T/\epsilon^2$ and $\beta_\epsilon = v_0\epsilon/(2D_T)$. Therefore, the diffusive and self-propulsion mechanisms in (17) are now accounted for together: jumping in the direction opposite to your orientation reduces the rate to $\sim \alpha_\epsilon(1 - \beta_\epsilon)$, whereas the there is a positive bias $\sim \alpha_\epsilon(1 + \beta_\epsilon)$ towards jumps in the direction pointed to by $\mathbf{e}(\theta)$, see Fig. 2. The tumbling (point 3 above) is replaced by a rotational Brownian motion. Taking the limit $\epsilon \to 0$ while keeping the occupied fraction $\phi = N\epsilon^2$ finite, one obtains the following macroscopic model for $f = f(\mathbf{x}, \theta, t)$:

$$\partial_t f + v_0 \nabla \cdot [f(1 - \phi\rho)\mathbf{e}(\theta)] = D_T \nabla \cdot ((1 - \phi\rho)\nabla f + \phi f \nabla \rho) + D_R \partial_\theta^2 f. \tag{22}$$

This model can be directly related to the fully discrete model (21): they are exactly the same if one considers (21) as the discretised-in-angle version of (22) by identifying

$$D_R \partial_\theta^2 f_k \approx D_R \frac{f_{k+1} + f_{k-1} - 2f_k}{(2\pi/m)^2},$$

that is, $\lambda = D_R m^2/(2\pi^2)$, where m is the number of orientations in the fully discrete model.

The other possible hybrid model is to consider a continuous random walk with interactions in space (2a), while only allowing a finite number of orientations, $\Theta_i \in$

$\{\theta_1, \dots, \theta_m\}$. In its simplest setting, we can consider that θ_k are equally spaced in $[0, 2\pi)$ and a constant switching rate λ between the neighbouring angles. The N particles evolve according to the stochastic model:

$$d\mathbf{X}_i = \sqrt{2D_T}d\mathbf{W}_i - \nabla_{\mathbf{x}_i} U dt + v_0 \mathbf{e}(\Theta_i)dt, \tag{23a}$$

$$\Theta_i = \{\theta_k\}_{k=1}^m, \quad \theta_k \xrightarrow{\lambda} \theta_{k+1} \pmod{2\pi}, \quad \theta_k \xrightarrow{\lambda} \theta_{k-1} \pmod{2\pi}. \tag{23b}$$

If we assume excluded-volume interactions through a hard-core potential, the resulting model is [67]

$$\partial_t f_k + v_0 \nabla \cdot [f_k(1-\phi\rho)\mathbf{e}_k + \phi \mathbf{p} f_k] = D_T \nabla \cdot [(1-\phi\rho)\nabla f_k + 3\phi f_k \nabla \rho] \\ + \lambda \left(f_{k+1} + f_{k-1} - 2f_k\right), \tag{24}$$

where $\rho = \sum_{k=1}^m f_k$, $\mathbf{p} = \sum_{k=1}^m f_k \mathbf{e}(\theta_k)$, and $\mathbf{e}_k = \mathbf{e}(\theta_k)$. The density $f_k(\mathbf{x}, t)$ represents the probability of finding a particle at position $\mathbf{x}$ at time t with orientation θ_k (naturally, we identify $f_{m+1} = f_1$ and $f_{-1} = f_m$). Here, $\phi = (N-1)\epsilon^2\pi/2$ represents the effective excluded region as in (15). We note how this model is consistent with the continuous model (15), in that if we had discretised angle in (15), we would arrive at the cross-diffusion reaction model (24).

A variant of the hybrid model (23) is to allow for jumps to arbitrary orientations instead of rotations of $2\pi/m$, namely, from θ_k to $\theta_j \pmod{2\pi}$, $j \neq k$, at a constant rate λ independent of the rotation. This is a convenient way to model the tumbles of a run-and-tumble process, such as the one used to describe the motion of *E. coli* [9], see also Sect. 6.2. In this case, the reaction term in (24) changes to

$$\partial_t f_k + v_0 \nabla \cdot [f_k(1-\phi\rho)\mathbf{e}(\theta_k) + \phi \mathbf{p} f_k] = D_T \nabla \cdot [(1-\phi\rho)\nabla f_k + 3\phi f_k \nabla \rho] \\ + \lambda \sum_{j \neq k} \left(f_j - f_k\right). \tag{25}$$

We may generalise the jumps in orientation by introducing a turning kernel $T(\theta, \theta')$ as the probability density function for a rotation from θ' to θ. That is, if $\Theta_i(t)$ is the orientation of the ith particle at time t and the jump occurs at t^*,

$$T(\theta, \theta')d\theta = \mathbb{P}\left(\left\{\theta \le \Theta_i(t_+^*) \le \theta + d\theta \mid \Theta_i(t_-^*) = \theta'\right\}\right).$$

Clearly, for mass conservation, we require that $\int T(\theta, \theta')d\theta = 1$. The jumps may only depend on the relative orientation $\theta - \theta'$ in the case of a homogeneous and isotropic medium, in which case $T(\theta, \theta') \equiv T(\theta - \theta')$. This is the case of the two particular examples above: in (24), the kernel is

$$T(\theta, \theta') = \frac{1}{2}\left[\delta(\theta - \theta' - \Delta) + \delta(\theta - \theta' + \Delta)\right], \qquad \Delta = \frac{2\pi}{m},$$

whereas the rotation kernel in (25) is

$$T(\theta, \theta') = \frac{1}{m-1} \sum_{k=1}^{m-1} \delta(\theta - \theta' + k\Delta), \qquad \Delta = \frac{2\pi}{m},$$

where the argument of the delta function is taken to be 2π-periodic. If the turning times t^* are distributed according to a Poisson process with intensity λ, the resulting macroscopic model for the phase density $f = f(\mathbf{x}, \theta, t)$ with a general turning kernel T becomes

$$\begin{aligned} \partial_t f + v_0 \nabla \cdot [f(1-\phi\rho)\mathbf{e}(\theta) + \phi \mathbf{p} f] &= D_T \nabla \cdot [(1-\phi\rho)\nabla f + 3\phi f \nabla \rho] \\ &- \lambda f + \lambda \int_0^{2\pi} T(\theta, \theta') f(\mathbf{x}, \theta', t) \mathrm{d}\theta'. \end{aligned} \tag{26}$$

We note that the microscopic process associated with (26) is continuous (and not hybrid) if the support of T has positive measure.

3 Models for Externally Activated Particles

In this section, we go from active to passive particles and consider models with time reversal at the microscopic level. As mentioned in the introduction, the defining factor of active matter models is the self-propulsion term, which makes them out of equilibrium. Mathematically, this can be expressed by saying that even the microscopic model lacks a gradient-flow structure (either due to the term $\mathbf{e}(\theta)$ in the transport term, see (15), or due to the reaction terms in (17) and (20), see Sect. 4).

In the previous section, we have seen the role the orientation θ plays. If it is kept continuous, the resulting macroscopic model is of kinetic type for the density $f(\mathbf{x}, \theta, t)$. If instead only a fixed number m of orientations are allowed, then these define a set of m species, whereby all the particles in the same species have the same drift term. This motivates the connection to cross-diffusion systems for passive particles, which are obtained by turning off the active change in directions in the models of Sect. 2 and look at the resulting special cases. This is a relevant limit in many applications, such as in pedestrian dynamics (see Sect. 6.1). Once the orientations are fixed, we are left with two possible passive systems: either originating from a spatial Brownian motion or from a spatial ASEP discrete process.

3.1 *Continuous Models*

The starting point is the microscopic model (2) taking the limit $D_R \to 0$. We could still keep the interaction potential as depending on the relative orientations, which would lead to different self- and cross-interactions (which might be useful in certain applications). Here, for simplicity, we assume interactions are all the same regardless of the orientations:

$$\mathrm{d}\mathbf{X}_i = \sqrt{2D_T}\mathrm{d}\mathbf{W}_i - \nabla_{\mathbf{x}_i} U \mathrm{d}t + v_0 \mathbf{e}(\Theta_i)\mathrm{d}t, \tag{27a}$$

$$\Theta_i(t) = \theta_k, \qquad \text{if } i \in \mathcal{I}_k, \qquad k = 1, \ldots, m, \tag{27b}$$

where $\mathcal{I}_k$ is the set of particles belonging to species k. The number of particles in each species is $|\mathcal{I}_k| = N_k$.

The mean-field limit of (27) is given by (taking $N = \sum_k N_k \to \infty$ as in (6a))

$$\partial_t f_k(\mathbf{x}, t) = \nabla_{\mathbf{x}} \cdot [D_T \nabla_{\mathbf{x}} f_k - v_0 \mathbf{e}(\theta_k) f_k + f_k \nabla_{\mathbf{x}}(u * \rho)], \tag{28}$$

and $\rho(\mathbf{x}, t) = \sum_k f_k$. For consistency with the active models, here, we do not take f_k to be probability densities but to integrate to the relative species fraction, whereas as before the total density ρ has unit mass:

$$\int_\Omega f_k(\mathbf{x}, t)\mathrm{d}\mathbf{x} = \frac{N_k}{N}, \qquad \int_\Omega \rho(\mathbf{x}, t)\mathrm{d}\mathbf{x} = 1. \tag{29}$$

Thus, $f_k = f_k(\mathbf{x}, t)$ describes the probability that a particle is at position $\mathbf{x}$ at time t *and* is in the $\mathcal{I}_k$ set.

The microscopic model (27) with the interaction term U replaced by a hard-core potential for particles with diameter ϵ can be dealt with via the method of matched asymptotics. In this case, the resulting cross-diffusion model is

$$\begin{aligned}\partial_t f_k + v_0 \nabla \cdot [f_k \mathbf{e}_k + \phi_{kl}(\mathbf{e}_l - \mathbf{e}_k) f_k f_l] = D_T \nabla \cdot [(1 + \phi_{kk} f_k) \nabla f_k \\ + \phi_{kl}(3 f_k \nabla f_l - f_l \nabla f_k)], \qquad l \neq k,\end{aligned} \tag{30}$$

where $\phi_{kk} = (N_k - 1)N/N_k \epsilon^2 \pi$, $\phi_{kl} = N\epsilon^2\pi/2$ for $l \neq k$ and $f_k(\mathbf{x}, t)$ are defined as above. This model was first derived in [14] for just two species but in a slightly more general context, whereby particles may have different sizes and diffusion coefficients (also, note that in [14], (30) appears written in terms of probability densities). Equation (30) can be directly related to model (24) with $\lambda = 0$ if in both models we assume N_k large enough such that $N_k - 1 \approx N_k$, $N - 1 \approx N$:

$$\partial_t f_k + v_0 \nabla \cdot [f_k(1 - \phi\rho)\mathbf{e}(\theta_k) + \phi \mathbf{p} f_k] = D_T \nabla \cdot [(1 - \phi\rho)\nabla f_k + 3\phi f_k \nabla \rho], \tag{31}$$

where $\phi = N\epsilon^2\pi/2$, $\rho = \sum_k f_k$, and $\mathbf{p} = \sum_k f_k \mathbf{e}(\theta_k)$. Model (31) is the cross-diffusion system for red and blue particles studied in [16] in disguise. First, set the number of species to $m = 2$ and define the number densities

$$r(\mathbf{x}, t) = N f_1(\mathbf{x}, t), \qquad b(\mathbf{x}, t) = N f_2(\mathbf{x}, t), \tag{32}$$

which integrate to N_1 and N_2, respectively. Then, define the potentials $V_r = -(v_0/D_T)\mathbf{e}(\theta_1) \cdot \mathbf{x}$ and $V_b = -(v_0/D_T)\mathbf{e}(\theta_2) \cdot \mathbf{x}$. In terms of these new quantities, system (31) becomes

$$\partial_t r = D_T \nabla \cdot [(1 + 2\varphi r - \varphi b)\nabla r + 3\varphi r \nabla b + r \nabla V_r + \varphi r b \nabla (V_b - V_r)]\,, \tag{33a}$$

$$\partial_t b = D_T \nabla \cdot [(1 + 2\varphi b - \varphi r)\nabla b + 3\varphi b \nabla r + b \nabla V_b + \varphi r b \nabla (V_r - V_b)]\,, \tag{33b}$$

where $\varphi = \epsilon^2 \pi/2$. This is exactly the cross-diffusion system for particles of the same size and diffusivity studied in [16] for $d = 2$ (see Eq. (11) in [16]).[1]

3.2 Discrete Models

In this category, there are discrete processes in space without changes in orientations. The most well-known model in the context of excluded-volume interactions is ASEP, which was used above in combination of either continuous change in angle, see (22), or discrete jumps, see (21). We obtain the corresponding passive process by setting either D_R or λ to zero, respectively. The resulting model in either case is

$$\partial_t f_k + v_0 \nabla \cdot [f_k(1-\phi\rho)\mathbf{e}(\theta_k)] = D_T \nabla \cdot [(1 - \phi\rho)\nabla f_k + \phi f_k \nabla \rho], \quad k = 1, \ldots, m, \tag{34}$$

where f_k satisfy (29) as before, and $\phi = N\epsilon^2$. We notice three differences with its continuous passive counterpart (31): in the latter, the effective occupied fraction ϕ has a factor of $\pi/2$, the coefficient in the cross-diffusion term $f_k\rho$ has a factor of three and the transport term has an additional nonlinearity that depends on the polarisation. The cross-diffusion system (34) was derived in [63] and analysed in [24] for two species ($m = 2$). Specifically, if we introduce the number densities r, b and general potentials V_r, V_b as above, it reads

$$\partial_t r = D_T \nabla \cdot [(1 - \bar\rho)\nabla r + r \nabla \bar\rho + r(1 - \bar\rho)\nabla V_r] \tag{35a}$$

$$\partial_t b = D_T \nabla \cdot [(1 - \bar\rho)\nabla b + b \nabla \bar\rho + b(1 - \bar\rho)\nabla V_b]\,, \tag{35b}$$

where $\bar\rho = \epsilon^2(r + b) = \epsilon^2(N_1 f_1 + N_2 f_2)$ (compare with (3.7)–(3.8) in [24]).[2]

[1] We note a typo in [16]: the coefficient β below system (11) should have read $\beta = (2d - 1)\gamma$.

[2] In the system (3.7)–(3.8) of [24], r and b are volume concentrations, thus having a factor of ϵ^2 compared to those used in (35), and the diffusivities of the two species are 1 and D instead of D_T for both.

4 General Model Structure

We now put the models presented in the previous sections into a more general picture. We assume that $f = f(\mathbf{x}, \theta, t)$, where θ is a continuous variable taking values in $[0, 2\pi)$ or a discrete variable taking values θ_k for $k = 1, \ldots, m$ (ordered increasingly on $[0, 2\pi)$). For now on, we consider the density rescaled by ϕ instead of a probability density. This implies that ϕ disappears from the equations and enters the mass condition as $\iint f = \int \rho = \phi$. In the latter case, we shall also use the notation $f_k(\mathbf{x}, t) = f(\mathbf{x}, \theta_k, t)$. We also recall the definition of the space density ρ and the polarisation $\mathbf{p}$:

$$\rho(\mathbf{x}, t) = \int_0^{2\pi} f(\mathbf{x}, \theta, t)\, \mathrm{d}\mu(\theta) \quad \text{and} \quad \mathbf{p}(\mathbf{x}, t) = \int_0^{2\pi} \mathbf{e}(\theta) f(\mathbf{x}, \theta, t)\, \mathrm{d}\mu(\theta),$$

where the integral in θ is either with respect to the Lebesgue measure for continuum angles or with respect to a discrete measure (a finite sum) for discrete angles.

The models presented have the following general model structure:

$$\partial_t f + v_0 \nabla \cdot (f(1-\rho)\mathbf{e}(\theta) + a\phi\mathbf{p}f) = D_T \nabla \cdot (\mathcal{B}_1(\rho)\nabla f + \mathcal{B}_2(f)\nabla\rho) + c\Delta_\theta f, \tag{36}$$

with $a \in \{0, 1\}$. In (36), the derivative operator ∇ is the standard gradient with respect to the spatial variable $\mathbf{x}$, while the Laplacian Δ_θ is either

- the second derivative $\partial_{\theta\theta} f$ in the Brownian case
- the second-order difference or discrete Laplacian

  $$\mathcal{D}^2 f = (f_{k+1} + f_{k-1} - 2f_k),$$

  with cyclic extension of the index k, in the case of fixed discrete rotations (or in one spatial dimension where there are only two possible orientations)
- the graph Laplacian with uniform weights

  $$\mathcal{D}_G f = \sum_{j \neq k} (f_j - f_k),$$

  in the run-and-tumble case (25) where arbitrary rotations are allowed.

Let us mention that similar structures and results hold true for graph Laplacians with other non-negative weights. We provide an overview of the respective differential operators and constants for most of the presented models in Table 1.

Small and Large Speed
Natural scaling limits for the general system (36) are the ones for small and large speed, i.e. $v_0 \to 0$ and $v_0 \to \infty$, respectively. The first case is rather obvious, since at $v_0 = 0$, the model is purely diffusive, i.e.

Table 1 Table recasting most models in the general form of (36)

Eq. No.	Δ_θ	a	$\mathcal{B}_1$	$\mathcal{B}_2$	c
(11)	$\partial_{\theta\theta}$	0	1	0	D_R
(15)	$\partial_{\theta\theta}$	1	$(1-\rho)$	$3f$	D_R
(20)	$\mathcal{D}^2$	0	1	0	λ
(21)	$\mathcal{D}^2$	0	$(1-\rho)$	f	λ
(22)	$\partial_{\theta\theta}$	0	$(1-\rho)$	f	D_R
(24)	$\mathcal{D}^2$	1	$(1-\rho)$	$3f$	D_R
(25)	$\mathcal{D}_G$	1	$(1-\rho)$	$3f$	D_R
(30)	None	0	$(1-\rho)$	$3f$	0
(34)	None	0	$(1-\rho)$	f	0

$$\partial_t f = D_T \nabla \cdot (\mathcal{B}_1(\rho)\nabla f + \mathcal{B}_2(f)\nabla\rho) + c\Delta_\theta f.$$

The model can then be written as a gradient-flow structure (or a generalised gradient structure in the case of discrete angles, see for example [50, 55]) for an entropy of the form

$$\mathcal{E}(f) = \iint f \log f \,\mathrm{d}\mathbf{x}\,\mathrm{d}\theta + b_2 \int (1-\rho)\log(1-\rho)\,\mathrm{d}\mathbf{x}, \tag{37}$$

with $b_2 \in \{0, 1, 3\}$ corresponding to the coefficients of $\mathcal{B}_2$. In the case v_0 small but finite, the gradient-flow structure is broken, but we still expect the diffusive part to dominate. In particular, we expect long-time convergence to a unique stationary solution.

In the case $v_0 \to \infty$, there are two relevant time scales. At a small time scale L/v_0, where L is a typical length scale, the evolution is governed by the first-order equation

$$\partial_\tau f + \nabla \cdot (f(1-\phi\rho)\mathbf{e}(\theta) + a\phi\mathbf{p}f) = 0,$$

where $\tau = tv_0/L$. The divergence of the corresponding velocity field $\mathbf{u} = (1 - \phi\rho)\mathbf{e}(\theta) + a\phi\mathbf{p}$ is given by

$$\nabla \cdot \mathbf{u} = -\phi\nabla\rho \cdot \mathbf{e}(\theta) + a\phi\nabla \cdot \mathbf{p}.$$

In particular, in the case of $a = 0$, we see that the question of expansion or compression of the velocity field is determined by the angle between $\nabla\rho$ and the unit vector $\mathbf{e}(\theta)$. Unless $\nabla\rho = 0$, the velocity field is compressible for a part of the directions and expansive for the opposite directions. A consequence to be expected is the appearance of patterns with almost piecewise constant densities (see, for example, Figs. 8 and 9). Inside the structures with constant densities ($\nabla\rho = 0$), the velocity field is incompressible, while the compression or expansion arises at the boundaries of such regions. This is rather described by a large time scale, i.e. the equation without time rescaling. Then, one expects a slow interface motion, which is also observed in numerical simulations. In a simple case with only one direction, this has been made precise in [18].

Small and Large Rotational Diffusion
The limit of small rotations rate $c \to 0$ corresponds to a more standard nonlinear Fokker–Planck system with a given linear potential,

$$\partial_t f + v_0 \nabla \cdot (f((1-\phi\rho)\mathbf{e}(\theta) + a\phi\mathbf{p})) = D_T \nabla \cdot (\mathcal{B}_1(\rho)\nabla f + \mathcal{B}_2(f)\nabla\rho),$$

as described in Sect. 3. Models of this kind have been investigated previously, see for example [24, 23]. They tend to develop patterns such as jams or lanes, depending on the initial condition. This happens in particular for large speeds v_0 (see Figs. 11 and 13).

The case of large rotational diffusion $c \to \infty$ will formally lead to f being constant with respect to θ at leading order. The corresponding equation at leading order can thus be obtained by averaging (36) in θ. Since f does not depend on θ, the polarisation is zero, that is

$$\mathbf{p} = \int_0^{2\pi} f\mathbf{e}(\theta)\,\mathrm{d}\theta = 0,$$

and the transport term drops out in all the models. Indeed, the nonlinear diffusion terms in any case average to linear diffusion with respect to $\mathbf{x}$. Hence, the evolution of f at leading order is governed by a linear diffusion equation.

4.1 Wasserstein Gradient Flows

We have seen above that microscopic models for externally activated particle have an underlying gradient-flow structure, which should ideally be maintained in the macroscopic limit. Adams et al. [1] showed in their seminal work that, then, the Wasserstein metric arises naturally in the mean-field limit (under suitable scaling assumptions). However, this limit is only well understood in a few cases (for example, for point particles) and rigorous results are often missing. In case of excluded-volume effects, as discussed in Sects. 2.1.1 and 2.3, the only known rigorous continuum models are derived in 1D [59, 11, 39], with only approximate models for higher space dimension. We see that these approximate limits often lack a full gradient-flow structure but are sufficiently close to it. In the following, we give a brief overview on how Wasserstein gradient flows and energy dissipation provide useful a priori estimates that can be used in existence proofs or when studying the long-time behaviour of solutions. These techniques are particularly useful for systems with cross-diffusion terms, for which standard existence results do not necessarily hold.

We will outline the main ideas for functions $f = f(\mathbf{x}, \theta, t)$, where θ is either continuous or taking discrete values θ_k with $k = 1, \dots m$. As before, we use $\xi = (\mathbf{x}, \theta)$. We say that a macroscopic model has a Wasserstein gradient-flow structure if it can be written as

$$\partial_t f(\mathbf{x}, \theta, t) = \nabla_\xi \cdot \big(\mathcal{M}(f) \nabla_\xi w\big), \tag{38}$$

where $\mathcal{M}$ is the mobility operator and $w = \delta_f \mathcal{E}$ the variational derivative of an entropy/energy functional $\mathcal{E}$ with respect to f. Note that for discrete θ_k, $k = 1, \dots m$, the mobility $\mathcal{M}$ is a positive definite matrix in $\mathbb{R}^{m \times m}$ and $\delta_f \mathcal{E}$ is replaced by the vector $\delta_{f_k} \mathcal{E}$. We have seen a possible candidate for energies in (37); they usually comprise negative logarithmic entropy terms of the particle distribution and the total density (corresponding to linear and nonlinear diffusion relating to the operators $\mathcal{B}_1$ and $\mathcal{B}_2$) as well as potentials.

If the system has a Wasserstein gradient-flow structure (38), then the entropy $\mathcal{E}$ changes in time as

$$\frac{d\mathcal{E}}{dt} = \iint \partial_t f w \, \mathrm{d}\mathbf{x}\mathrm{d}\theta = -\iint \bar{\mathcal{M}}(w) |\nabla_\xi w|^2 \, \mathrm{d}\mathbf{x}\mathrm{d}\theta, \tag{39}$$

where $\bar{\mathcal{M}}$ is the mobility matrix $\mathcal{M}$ written in terms of the entropy variable w. If $\bar{\mathcal{M}}$ is positive definite, then the energy is dissipated. In the next subsection, we will define an entropy for the general model (36) and show that the system is dissipative for several of the operator choices listed in Table 1.

Note that these entropy dissipation arguments are mostly restricted to unbounded domains and bounded domains with no-flux or Dirichlet boundary conditions. It is possible to generalise them in the case of non-equilibrium boundary conditions, as such discussed in Sect. 5, but a general theory is not available yet. We will see in the next subsection that entropy dissipation may also hold for systems, which do not have a full gradient-flow structure.

Since system (38) is dissipative, we expect long-time convergence to an equilibrium solution. The respective equilibrium solutions f_∞ to (38) then correspond to minimisers of the entropy $\mathcal{E}$. To show exponential convergence towards equilibrium, it is often helpful to study the evolution of the so-called relative entropy, that is,

$$\mathcal{RE}(f, f_\infty) := \mathcal{E}(f) - \mathcal{E}(f_\infty) - \langle \mathcal{E}'(f_\infty), f - f_\infty \rangle.$$

In general, one wishes to establish the so-called entropy–entropy dissipation inequalities for the relative entropy

$$\frac{d\mathcal{RE}}{dt} \leq -C\mathcal{RE},$$

with $C > 0$. Then, Gronwall's lemma gives desired exponential convergence. This approach is also known as the Bakry–Emery method, see [6].

We discussed the challenges in the rigorous derivation of continuum models in the previous sections and how often only formal or approximate limiting results are available. These approximate models are often 'close' to a full gradient flow, meaning that they only differ by higher order terms (which were neglected in the approximation). This closeness motivated the definition of so-called *asymptotic gradient flow*, see [16, 15]. A dynamical system of the form

$$\partial_t z = \mathcal{F}(z;\epsilon) \tag{40}$$

has an asymptotic gradient-flow structure of order k if

$$\mathcal{F}(z;\epsilon) + \sum_{j=k+1}^{2k} \epsilon^j \mathcal{G}_j(z) = -\mathcal{M}(z;\epsilon)\mathcal{E}'(z,\epsilon),$$

for some parametric energy functional $\mathcal{E}$. For example, (30) exhibits a GF structure if the red and blue particles have the same size and diffusivity but lacks it for differently sized particles (a variation of the model not discussed here). The closeness of AGF to GF can be used to study, for example, its stationary solutions and the behaviour of solution close to equilibrium, see [3, 2, 16].

4.2 Entropy Dissipation

Next, we investigate the (approximate) dissipation of an appropriate energy for the general formulation (36). The considered energy functional is motivated by the entropies of the scaling limits considered before. In particular, we consider

$$\mathcal{E}(f) = \iint f \log f + V(\mathbf{x},\theta) f \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mu(\theta) + b_2 \int (1-\rho)\log(1-\rho)\,\mathrm{d}\mathbf{x}, \tag{41}$$

for which the models can be formulated as gradient flows in the case $c = 0$ (no active self-propulsion) with $b_2 \in \{0, 1, 3\}$ chosen appropriately. For simplicity, we set $\phi = 1$ as well as $D_T = 1$ in the following. As before, we interpret integrals in θ with respect to the Lebesgue measure for continuum angles and with respect to the discrete measure (sum) in case of a finite number of directions. We recall that the potential V is given by

$$V(\mathbf{x},\theta) = -v_0\,\mathbf{e}(\theta)\cdot\mathbf{x} = -v_0\,(\cos\theta x + \sin\theta y).$$

In the following, we provide a formal computation assuming sufficient regularity of all solutions. We have

$$
\begin{aligned}
\frac{d\mathcal{E}}{dt} &= \iint \partial_t f(\log f + V - b_2 \log(1-\rho))\, \mathrm{d}\mathbf{x}\, \mathrm{d}\theta \\
&= -\iint \nabla \left[\log f + V - b_2 \log(1-\rho)\right] \{-v_0 f[(1-\rho)\mathbf{e}(\theta) + a\mathbf{p}] \\
&\quad + \mathcal{B}_1(f,\rho)\nabla f + \mathcal{B}_2(f,\rho)\nabla\rho\}\ \mathrm{d}\mathbf{x}\, \mathrm{d}\theta \\
&\quad + c \iint (\log f + V - b_2 \log(1-\rho))\Delta_\theta f\ \mathrm{d}\mathbf{x}\, \mathrm{d}\theta.
\end{aligned}
$$

Let us first investigate the last term. Since ρ is independent of θ, using the properties of the generalised Laplacian Δ_θ with periodic boundary conditions, we have

$$
\int \log(1-\rho)\Delta_\theta f\, d\theta = \log(1-\rho) \int \Delta_\theta f\, d\theta = 0.
$$

Using the fact that $\Delta_\theta \mathbf{e}(\theta)$ is uniformly bounded in all cases, we find

$$
\begin{aligned}
\iint [\log f + V - b_2 \log(1-\rho)]\Delta_\theta f\ \mathrm{d}\mathbf{x}\, \mathrm{d}\theta &= -\iint \mathcal{F}_\theta(f) - v_0 \Delta_\theta \mathbf{e}(\theta)\cdot \mathbf{x} f\ d\mathbf{x}\, d\theta, \\
&\le C|v_0| \int |\mathbf{x}| f\ \mathrm{d}\mathbf{x}\, \mathrm{d}\theta = C|v_0| \int |\mathbf{x}|\rho\ \mathrm{d}\mathbf{x},
\end{aligned}
$$

where $\mathcal{F}_\theta(f) \ge 0$ is the Fisher information with respect to the generalised Laplacian Δ_θ

$$
\mathcal{F}_\theta(f) = \begin{cases}
\dfrac{|\partial_\theta f|^2}{f} & \text{for } \Delta_\theta = \partial_{\theta\theta}, \\
\dfrac{|f_{k+1} - f_k|^2}{M(f_k, f_{k+1})} & \text{for } \Delta_\theta = \mathcal{D}^2, \\
\displaystyle\sum_j \frac{|f_j - f_k|^2}{M(f_j, f_k)} & \text{for } \Delta_\theta = \mathcal{D}_G,
\end{cases}
$$

where

$$
M(f,g) = \frac{f-g}{\log(f) - \log(g)}
$$

is the logarithmic mean.

Now, we further investigate the first term for the models with $a = 0$ (no $\mathbf{p}$ term in the equation for f), where, for the respective b_2, we obtain

$$
\begin{aligned}
&\iint \nabla[\log f + V - b_2 \log(1-\rho)]\,[v_0 f(1-\rho)\mathbf{e}(\theta) - \mathcal{B}_1(f,\rho)\nabla f - \mathcal{B}_2(f,\rho)\nabla\rho]\ \mathrm{d}\mathbf{x}\, \mathrm{d}\theta \\
&\qquad = -\iint f(1-\rho)|\nabla[\log f + V - b_2 \log(1-\rho)]|^2\ \mathrm{d}\mathbf{x}\, \mathrm{d}\theta \le 0.
\end{aligned}
$$

Overall, we finally find

$$\frac{d\mathcal{E}}{dt} \leq C\ |v_0| \int |\mathbf{x}|\rho\ \mathrm{d}\mathbf{x} \leq C\ |v_0| \sqrt{\int |\mathbf{x}|^2\rho\ \mathrm{d}\mathbf{x}}.$$

Thus, the growth of the entropy in time is limited by the second moment. Note that for $a = 1$, one can employ analogous reasoning to obtain the above negative term. However, it is unclear how to control the additional term $\iint \nabla[\log f + V - c\log(1-\rho)]v_0\mathbf{p}f\ \mathrm{d}\mathbf{x}\ \mathrm{d}\theta$. The bounds obtained provide useful a priori estimates, which can be used in existence results, and to study the long-time behaviour, see for example [21, 45].

5 Boundary Effects

So far, we have focused on domains with periodic boundary conditions. In this section, we discuss non-zero-flux boundary conditions, which can be used to impose non-zero steady currents and externally activate or force the passive models described in Sect. 3 out of equilibrium [27]. We remark that the in-flux boundary conditions are difficult to deal with in the case of interacting continuous random walks. Thus, we only mention a few aspects and comment in more detail on the time-discrete situation which is easier to tackle, see Remark 5.1.

5.1 Mass Conserving Boundary Conditions

We first discuss conditions (other than periodic boundaries) that conserve the total mass, i.e. the total number of particles in the microscopic models, or the integral of the density ϕ in the macroscopic models. In case of the coupled SDE model (2), we are interested in conditions that ensure that particles remain inside the domain. Intuitively, particles need to be reflected whenever they hit the boundary. However, as we are dealing with a problem that is continuous in time, we have to ensure that the particle path remains continuous. In his seminal paper [64], Skorokhod solved this problem by introducing an additional process that increases whenever the original process hits the boundary, see [57] for a detailed discussion. For the microscopic models on a lattice, such boundary conditions correspond to aborting any jumps that would lead a particle outside of the domain. For the macroscopic models, mass conservation corresponds to no-flux boundary conditions that are implemented by setting the normal flux over the boundary to zero, i.e.

$$\mathbf{J} \cdot \mathbf{n} = 0 \text{ a.e. in } \Upsilon \times (0, T), \tag{42}$$

where, using the general form (36), the flux density is given as

$$\mathbf{J} = v_0(f((1-\phi\rho)\mathbf{e}(\theta) + a\phi\mathbf{p})) - D_T(\mathcal{B}_1(\rho)\nabla f + \mathcal{B}_2(f)\nabla\rho). \tag{43}$$

5.2 *Flux Boundary Conditions*

Apart from periodic or no-flux boundary conditions, there is also the possibility for boundary conditions that allow for the in- or outflow of particles (mass) via the boundary. Such effects are of particular interest in the context of this chapter, since they yield an out-of-equilibrium system even if the motion of the particles within the domain is purely passive (i.e. due to diffusion).

For the SDE model (2), such boundary conditions correspond to partially reflecting or radiation conditions. Intuitively, once a particle reaches the boundary it is, with a certain probability, either removed or otherwise reflected, see [41] and [49, Section 4]. For the discrete models of Sect. 2.2, let us consider first the special case of a single species in two dimensions with two open and two closed boundaries. This corresponds to the asymmetric simple exclusion process (ASEP) with open boundary conditions, the paradigmatic models in non-equilibrium thermodynamics, [28]. The dynamics of such a process is well understood and can be solved explicitly [33, 32] (see also [68]). We denote by α and β the rates by which particles enter (at the left boundary) and exit (at the right boundary) the lattice. Then, the key observation here is that in the steady state, system can be in one of the three distinct states, characterised by the value of the one-dimensional current and the density as follows:

- *Low density* or *influx limited* ($\alpha < \min\{\beta, 1/2\}$): the density takes the value α and the flux $\alpha(1-\alpha)$.
- *High density* or *outflux limited* ($\beta < \min\{\alpha, 1/2\}$): the density is $1-\beta$ and the flux $\beta(1-\beta)$.
- *Maximal current* ($\alpha, \beta > 1/2$): the density is $1/2$ and the flux $1/4$.

A similar behaviour can be verified for the macroscopic passive model (35) (or also (19) with $\lambda = 0$) for a single species on the domain $\Omega = [0, L]$, which reduces to a single equation for the unknown density r, i.e.

$$\partial_t r + \partial_x j = 0 \text{ with } j = -D_T\partial_x r + r(1-r)\partial_x V.$$

We supplement the equation with the flux boundary conditions

$$-j\cdot n = \alpha(1-r) \text{ at } x = 0 \text{ and } j\cdot n = \beta r \text{ at } x = 1, \tag{44}$$

see [20]. Indeed, one can show that for positive $D_T > 0$, stationary solutions are close to one of the regimes, and as $D_T \to 0$, these attain the exact values for flux and density. Interestingly, for positive D_T, it is possible to enter the maximal current regime for values of α and β strictly less than $1/2$. The long-time behaviour of these equations, using entropy–entropy dissipation inequalities, has been studied in [20].

For the macroscopic active models (15) and (22), a similar condition can be formulated for the unknown quantity f. However, as f depends not only on $\mathbf{x}$ and t but also on the angle θ, the coefficients may also depend on it. In the most general situation, we obtain

$$\mathbf{J} \cdot \mathbf{n} = -\alpha(\theta, \mathbf{n})(1 - \phi\rho) + \beta(\theta, \mathbf{n}) f, \tag{45}$$

with $\mathbf{J}$ defined in (43). Here, the choice of the functions α and β is subject to modelling assumptions or properties of microscopic stochastic models for the inflow and outflow. Typically, one has a separation into inflow and outflow regions, which means that α is supported on inward pointing directions $\mathbf{e}(\theta) \cdot \mathbf{n} > 0$, while β is supported outward pointing directions $\mathbf{e}(\theta) \cdot \mathbf{n} > 0$.

5.3 *Other Boundary Conditions*

Let us also discuss other types of boundary conditions. Homogeneous Dirichlet boundary conditions can be applied to all types of models: for the SDE (2), one has to remove a particle once it reaches the boundary. The same holds for the discrete random walk models. For the macroscopic models, one sets the trace at the boundary to zero. Finally, also mixed boundary conditions are possible, combining the effects described above on different parts of the boundary. Another type of boundary condition useful in the context of self-propelled organisms is no-slip or alignment type boundary conditions, whereby the particles align their orientations with the boundary ($\mathbf{e}(\theta) \cdot \mathbf{n} = 0$). A notable example of this can be seen in ant foraging networks and lab experiments with ants walking on bridges [36, 35].

Remark 5.1 (Boundary Conditions for Discrete-Time Random Walks) We briefly comment on the situation for time-discrete random walks, that is, when the SDE (2) is replaced by the time-discrete system

$$\mathbf{X}_i(t + \Delta t) = \mathbf{X}_i(t) + \Delta t\sqrt{2D_T}\zeta_i - \Delta t \nabla_{\mathbf{x}_i} U + \Delta t v_0 \mathbf{e}(\Theta_i), \tag{46a}$$

$$\Theta_i(t + \Delta t) = \Theta_i(t) + \Delta t\sqrt{2D_R}\bar{\zeta}_i - \Delta t \partial_{\theta_i} U, \tag{46b}$$

for some time step size $\Delta t > 0$ and where ζ_i and $\bar{\zeta}_i$ are normally distributed random variables with zero mean and unit variance. To implement boundary conditions, one has to calculate the probability that $\mathbf{X}_i(t + \Delta t) \notin \Omega$ (considering also the case that the particles leave the domain but move back into it within the time interval $[t, t+\Delta t]$), see [5] for detailed calculations in the case of pure diffusion. If a particle is found to have left the domain, it can either be removed with probability one (corresponding to homogeneous Dirichlet boundary conditions) or less than one, called a partially reflective boundary condition (corresponding to Robin boundary conditions). In our setting, this probability can depend on the current angle of the particle, $\Theta_i(t)$, allowing for additional modelling. It is also possible to add a reservoir of particles at the boundary to implement flux boundary conditions in the spirit of (45) by prescribing a probability to enter the domain. In the case of excluded volume, the probability to enter will depend on the number of particles close to the entrance.

6 Active Crowds in the Life and Social Science

6.1 Pedestrian Dynamics

A prominent example of active and externally activated dynamics in the context of socio-economic applications is the motion of large pedestrian crowds. There is an extensive literature on mathematical modelling for pedestrians in the physics and the transportation community, which is beyond the scope of this chapter. We will therefore review the relevant models in the context of active crowds only and refer for a more comprehensive overview to [30, 51].

Microscopic Models for Pedestrian Flows
Microscopic off-lattice models are the most popular approach in the engineering and transportation research literature. Most software packages and simulations are based on the so-called social force model by Helbing [43, 42]. The social force model is a second-order SDE model, which does not take the form of active models considered here. However, it is easy to formulate models for pedestrians in the context of active particles satisfying (2). For example, assume that all pedestrians move with the same constant speed in a desired direction Θ_d avoiding collisions with others. Then, their dynamics can be described by the following second-order system:

$$\mathrm{d}\mathbf{X}_i = -\nabla_{\mathbf{X}_i} U \mathrm{d}t + v_0 \frac{e(\Theta_i) - \Theta_d}{\tau} dt + \sqrt{2D_T}\, \mathrm{d}\mathbf{W}_i \tag{47a}$$

$$\mathrm{d}\Theta_i = -\partial_{\Theta_i} U \mathrm{d}t + \sqrt{2D_R}\, \mathrm{d}\mathbf{W}_i . \tag{47b}$$

The potential U takes the form (1), where the pairwise interactions u should be related to the likelihood of a collision. One could for example consider

$$u(|\mathbf{X}_i - \mathbf{X}_j|/\ell, \Theta_i - \Theta_j) = C \frac{\Theta_i - \Theta_j}{|\mathbf{X}_i - \mathbf{X}_j|},$$

where $C \in \mathbb{R}^+$ and ℓ relates to the personal comfort zone. Another possibility corresponds to a Lennard-Jones type potential to model short-range repulsion and long-range attraction. Another popular microscopic approach is the so-called cellular automata, which correspond to the discrete active and externally activated models discussed before. In cellular automata, a certain number of pedestrians can occupy discrete lattice sites and individuals move to available (not fully occupied) neighbouring sites, with transition rates. These transition rates may depend on given potentials, as discussed in the previous sections, which relate to the preferred direction.

There is also a large class of microscopic on-lattice models, so-called cellular automata, see [47], which relate to the microscopic discussed in Sect. 3.2. In cellular automata, pedestrians move to neighbouring sites at given rates, if these sites are not already occupied. Their rates often depend on an external given potential, which relates to the desired direction Θ_d. Cellular automata often serve as the basis for the macroscopic pedestrian models, which will be discussed in the next paragraph, see for example [19, 22].

Macroscopic Models for Pedestrian Flows

Mean-field models derived from microscopic off-lattice approaches have been used successfully to analyse the formation of directional lanes or aggregates in bi-directional pedestrian flows. This segregation behaviour has been observed in many experimental and real-life situations. Several models, which fall into the category of externally activated particles introduced in Sect. 3.2, were proposed and investigated in this context. These models take the form (35), in which the densities r and b relate to different directions of motion. For example, in the case of bi-directional flows in a straight corridor, 'red particles' correspond to individuals moving to the right, while blue ones move to the left. We will see in Sect. 7.1 that we can observe temporal as well as stationary segregated states. Depending on the initial and inflow conditions, directional lanes or jams occur. Then, the gradient-flow structure can then be used to investigate the stability of stationary states using for example the respective entropy functionals. Due to the segregated structure of stationary solutions, one can also use linear stability analysis around constant steady states to understand for example the formation of lanes, see [56].

More pronounced segregated states and lanes can be observed when allowing for side-stepping. In the respective microscopic on lattice models, individuals step aside when approached by a member of the other species. The respective formally derived mean-field model has a perturbed gradient-flow structure, which can be used to show existence of solutions, see [22]. More recently, a model containing both an active and a passive species has been studied in [29].

6.2 Transport in Biological Systems

Another example where active particles play an important role is transport process in biological systems. We will discuss two important types of such processes in the following: chemotaxis and transport in neurones.

Chemotaxis

We consider bacteria in a given domain that aim to move along the gradient of a given chemical substance, called chemoattractant and modelled by a function $c : \Omega \to \mathbb{R}_+$. Due to their size, bacteria cannot sense a gradient by, say, comparing the value of c at their head with that at their tail. Thus, they use a different mechanism based on comparing values of c at different time instances and different points in space, called run and tumble. In a first step, they perform a directed motion into a fixed direction (run) and then rotate randomly (tumble). These two steps are repeated, and however, the probability of tumbling depends on c as follows: if the value of c is decreasing in time, bacteria tumble more frequently as they are not moving up the gradient. If the value of c increased, they turn less often. Roughly speaking, this mechanism reduces the amount of diffusion depending on the gradient of c. Here, we consider a slightly different idea that fits into the hybrid random walk model introduced in (23), assuming D_T to be small (run), and the rate of change for the angle depends on c. To this end, λ is taken different for each angle (thus denoted by λ_k) and is assumed to depend on the difference of the external signal c at the current and past positions, only. Denoting by t_k, $k = 1, 2, \ldots$, the times at which the angle changes, at time t_n, we have $\lambda_k = \lambda_k((c(\mathbf{X}_i(t_n)) - c(\mathbf{X}_i(t_{n-1}))$. Additionally, we introduce a fixed baseline turning frequency $\bar{\lambda}$ and consider

$$\lambda_k = \bar{\lambda} + (c(\mathbf{X}_k(t_{n-1})) - c(\mathbf{X}_k(t_n))).$$

Now, going from discrete to time-continuous jumps, i.e. $t_n - t_{n-1} \to 0$, and appropriate rescaling, we obtain via the chain rule

$$\lambda_k = \bar{\lambda} - \dot{\mathbf{X}}_k \cdot \nabla c(\mathbf{X}_i).$$

However, due to the stochastic nature of the equation governing the evolution $\mathbf{X}_k$, its time derivative is not defined. Thus, as a modelling choice, we replace this velocity vector by $v_0 \mathbf{e}(\theta_k)$, i.e. the direction of the active motion of the respective particle. This is also motivated by the fact that for $D_T = 0$ and $U = 0$ in (23a), this is exact. We obtain

$$\lambda_k = \bar{\lambda} - v_0 \mathbf{e}(\theta_k) \cdot \nabla c(\mathbf{X}_i).$$

In the particular case on one spatial dimension with only two possible angles (denoted by $+$ and $-$) and for $v_0 = 1$, this reduces to

$$\lambda_\pm = \bar{\lambda} \mp \partial_x c,$$

which is exactly the model analysed in [58]. There, it was also shown that using an appropriate parabolic scaling, one can obtain a Chemotaxis-like model with linear transport but nonlinear diffusion in the diffusive limit.

Transport in Neurones

Another interesting example is transport processes within cells, and we focus on the example of vesicles in neurones. Vesicles are small bubbles of cell membrane that are produced in the cell body (soma) and are then transported along extensions of the cell called axons. The transport itself is carried out by motor proteins that move along microtubules and are allowed to change their direction of motion. This situation can be modelled using the discrete random walks from Sect. 2.2 by considering the one-dimensional case which, in the macroscopic limit, yields Eq. (17). Since we are now dealing with two species f_- and f_+, denoting left- and right-moving complexes, we also have to adopt our boundary conditions as follows: denoting by j_+ and j_- the respective fluxes,

$$\begin{aligned} -j_+ &= \alpha_+(1-\phi\rho), \quad j_- = \beta_- f_- \qquad \text{at } x = 0, \\ -j_- &= \alpha_-(1-\phi\rho), \quad j_+ = \beta_+ f_+ \qquad \text{at } x = 1. \end{aligned}$$

System (17) has, to the best of our knowledge, not yet been considered with these boundary conditions. From an application point of view, it is relevant to study whether these models are able to reproduce the almost uniform distribution of motor complexes observed in experiments, see [13, 12] for an analysis.

More recently, the influence of transport in developing neurites has been studied in [44] with an emphasis on the mechanism that decides which of the growing neurites becomes an axon. To model this situation, the concentration of vesicles at some and growth cones is modelled separately by ordinary differential equations which are connected to instances of (35) via flux boundary conditions.

7 Numerical Simulations

7.1 One Spatial Dimension

In the following, we present numerical examples in one spatial dimension comparing a subset of models presented above. All simulations in this subsection are based on a finite element discretisation in space (using P1 elements). The time discretisation is based on the following implicit–explicit (IMEX) scheme:

$$\begin{aligned} \frac{f^{n+1}-f^n}{\tau} + v_0\nabla\cdot[f(1-\phi\rho^n)\mathbf{e}(\theta) + a\phi\mathbf{p}f] = D_T\nabla\cdot(\mathcal{B}_1(\rho^n)\nabla f^{n+1} \\ + \mathcal{B}_2(f^n)\nabla\rho^{n+1}) + c\Delta_\theta f^n, \end{aligned}$$

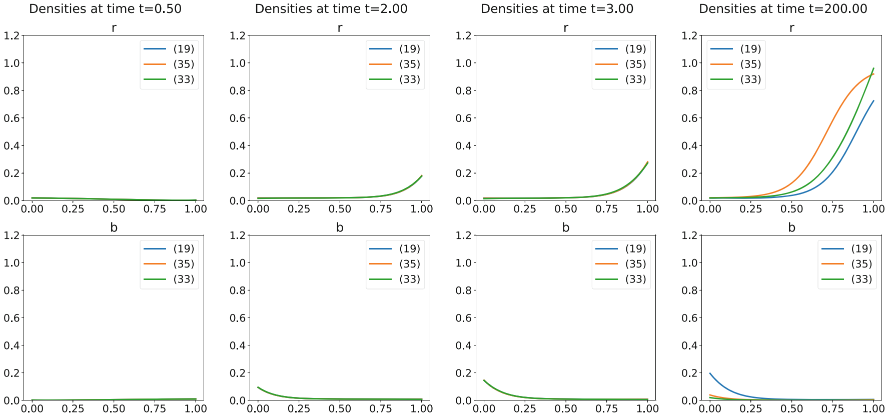
in which the superscript index n refers to the nth time step, that is, $t^n = n\tau$, $\tau > 0$. Here, transport and rotational diffusion are taken explicitly, while in the diffusive part terms of second order are treated implicitly. Thus, in every time step, a linear system has to be solved. All schemes were implemented using the finite element library NgSolve, see [62].

We will illustrate the behaviour of solutions for models (19), (33), (35), in case of in- and outflux (44), no-flux (42) or periodic boundary conditions in case of two species, referred to as red r and blue b particles. We use subscript r and b, when referring to their respective inflow and outflow rates as well as diffusion coefficients. Note that while for the models (19) and (35), the one-dimensional setting is meaningful, for model (33), the simulations are to be understood as two-dimensional but with a potential that is constant in the second dimension. For all simulations, we discretised the unit interval into 150 elements and chose time steps of size $\tau = 0.01$.

Flux Boundary Conditions

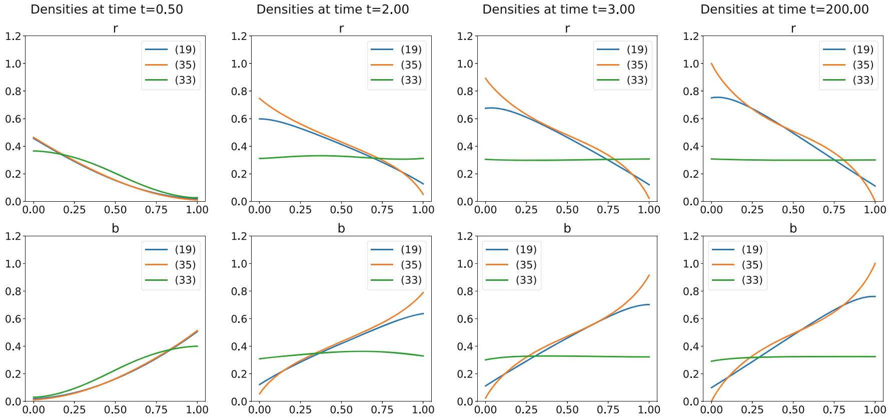
Figures 3 and 4 show density profiles for the respective models at time $t = 0.5, 2, 3, 30$. In Fig. 3, we chose rather low rates (in particular below $1/2$) and with $\alpha_r > \beta_r$ as well as $\alpha_b < \beta_b$ which resulted in species r being in a outflux limited and species b an in-flux limited phase. We observe that for these low rates, all models are quite close to one another, yet with different shapes of the boundary layers. The slope of solutions at the boundary seems to be influenced by the respective diffusion terms, in particular the cross-diffusion terms in (33) and (35).

In Fig. 4, we chose rates above $1/2$ to obtain the maximal current phase. There, interestingly, it turns out that the dynamics of model (33) shows a completely different behaviour. This constitutes an interesting starting point for further analytical considerations on the phase behaviour. Figure 5 displays the evolution of the total mass of the respective species for different inflow and outflow rates. We observe that the reaction–diffusion (19) and the lattice-based cross-diffusion system (35)



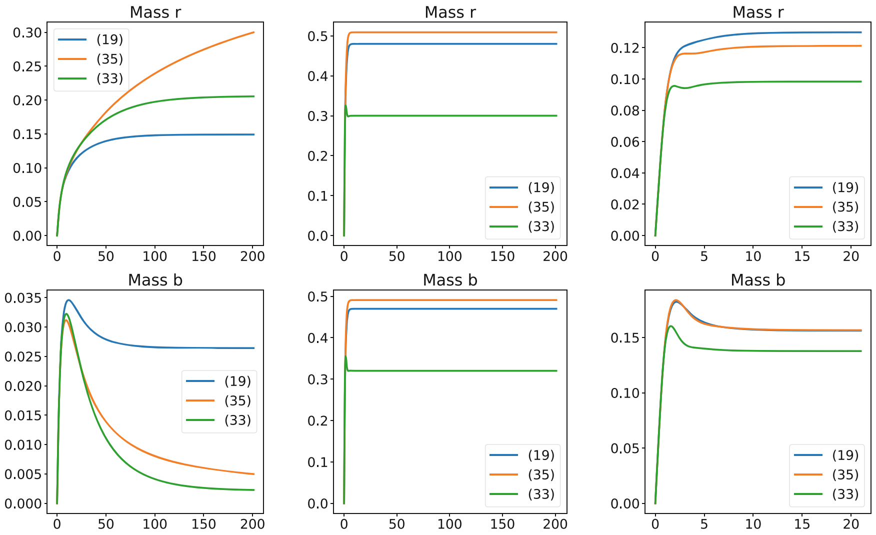
(a) (b) (c) (d)

Fig. 3 Flux boundary conditions with $\lambda = 0.01$, $D_r = 0.1$, $D_b = 0.1$, $\alpha_r = 0.02$, $\beta_r = 0.01$, $\alpha_b = 0.01$ and $\beta_b = 0.02$, which yields the in-flux-limited phase for species r and outflux-limited for b. (**a**) $t = 0.5$. (**b**) $t = 2$. (**c**) $t = 3$. (**d**) $t = 200$



(a) (b) (c) (d)

Fig. 4 Flux boundary conditions with $\lambda = 0.01$, $D_r = 0.1$, $D_b = 0.1$, $\alpha_r = 0.6$, $\beta_r = 0.8$, $\alpha_b = 0.7$, $\beta_b = 0.9$ which yields the maximal current phase. (**a**) $t = 0.5$. (**b**) $t = 2$. (**c**) $t = 3$. (**d**) $t = 200$



(a) (b) (c)

Fig. 5 Evolution of the total mass for different flux boundary conditions and with $D_r = D_b = 0.1$ and $\lambda = 0.01$ in all cases. (**a**) $\alpha_r = 0.02$, $\beta_r = 0.01$, $\alpha_b = 0.01$, $\beta_b = 0.02$. (**b**) $\alpha_r = 0.6$, $\beta_r = 0.8$, $\alpha_b = 0.7$, $\beta_b = 0.9$. (**c**) $\alpha_r = 0.1$, $\beta_r = 0.2$, $\alpha_b = 0.2$, $\quad \beta_b = 0.4$

show a similar quantitative behaviour in several inflow and outflow regimes, while the cross-diffusion system obtained via asymptotic expansion (33) behaves only qualitatively similar.

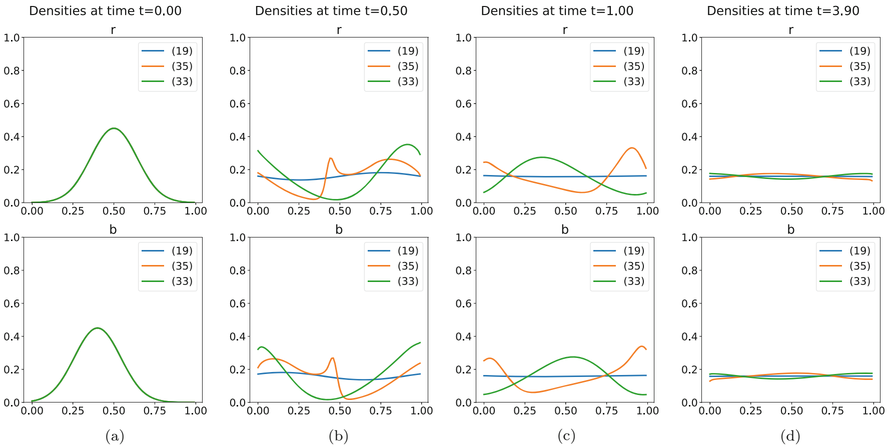


Fig. 6 Periodic boundary conditions with $D_r = D_b = 0.01$ and $\lambda = 0.01$. All models converge to constant stationary solution. (**a**) $t = 0$. (**b**) $t = 0.4$. (**c**) $t = 1$. (**d**) $t = 3.9$

Periodic Boundary Conditions

For periodic boundary conditions, noting that the velocity is constant, thus periodic, we expect constant stationary solutions whose value is determined by the initial mass. This is indeed observed in Fig. 6. However, for earlier times, their dynamics differs substantially, in particular for (35), the influence of cross-diffusion ('jams') is most pronounced.

Confining Potential

Finally, in Fig. 7, we consider the situation of no-flux conditions together with a confining potential $V(x) = (x - \frac{1}{2})^2$. Here, we observe very similar behaviour for all models, probably due to the fact that the transport term dominates the dynamics.

7.2 Two Spatial Dimensions

In this subsection, we reproduce numerical examples in two spatial dimensions from [17]. In particular, we show examples of the active continuous model (15), the active hybrid model (22) and the passive version of the latter (35), which corresponds to setting $D_R = 0$ in (22) and choosing an initial condition in angle of the form $\delta(\theta - \theta_1) + \delta(\theta - \theta_2)$ with θ_i such that $V_r = -(v_0/D_T)\mathbf{e}(\theta_1) \cdot \mathbf{x}$ and $V_b = -(v_0/D_T)\mathbf{e}(\theta_2) \cdot \mathbf{x}$. Throughout this subsection, we use periodic boundary conditions in the spatial domain $\Omega = [0, 1]^2$, as well as in the angular domain $[0, 2\pi]$ for the active models (15) and (22). We use the first-order finite-volume scheme of [17], which is based on [26, 61]. The scheme is implemented in Julia.

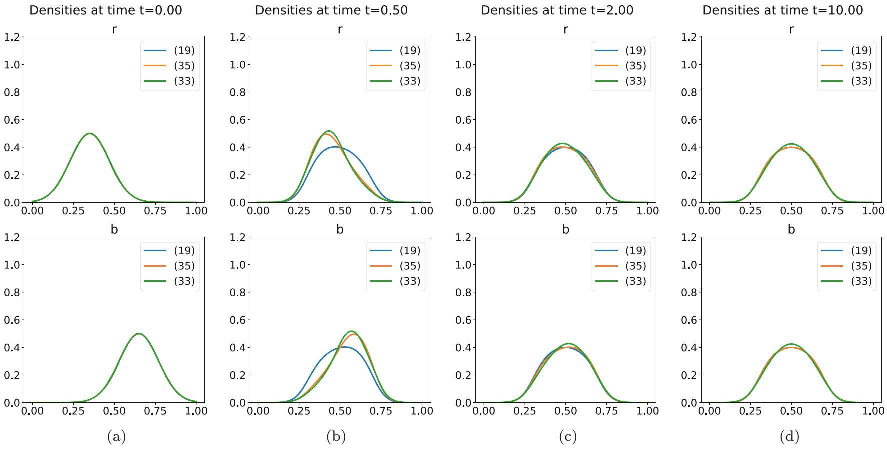


Fig. 7 No-flux boundary conditions with $D_r = D_b = 0.1$, $\lambda = 0.01$ and a confining potential $V_r = V_b = 5(x - 0.5)^2$. (**a**) $t = 0$. (**b**) $t = 0.5$. (**c**) $t = 2$. (**d**) $t = 10$

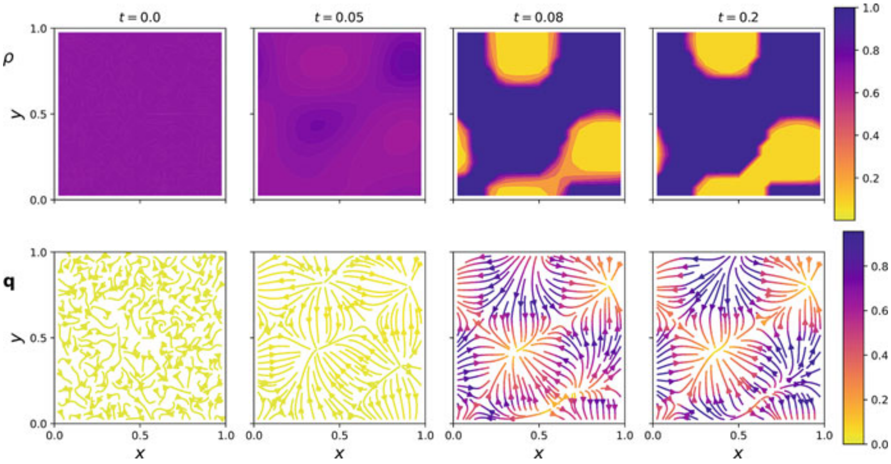


Fig. 8 Hybrid model for active particles (22) with periodic boundary conditions in $[0, 1]^2 \times [0, 2\pi]$ at different times, starting from a 3D random perturbation around the homogeneous solution. The first row shows the total density $\rho(\mathbf{x}, t)$ with mass ϕ and the second row the mean direction $(\mathbf{x}, t)$ (48). Parameters used: $D_T = D_R = 1$, $v_0 = 60$, $\phi = 0.7$

We use a discretisation with 21 points in each direction and a time step $\Delta t \leq 10^{-5}$ satisfying the CFL condition given by Theorem 3.2 in [26].

Figures 8 and 9 show the outputs of the two active models (15) and (22) using the same parameters, $D_T = D_R = 1$, $v_0 = 60$, $\phi = 0.7$. In both case, we observe the formation of *motility-induced phased separation* (MIPS), namely a separation into dilute and dense regions and a polarisation of particles in the boundary between

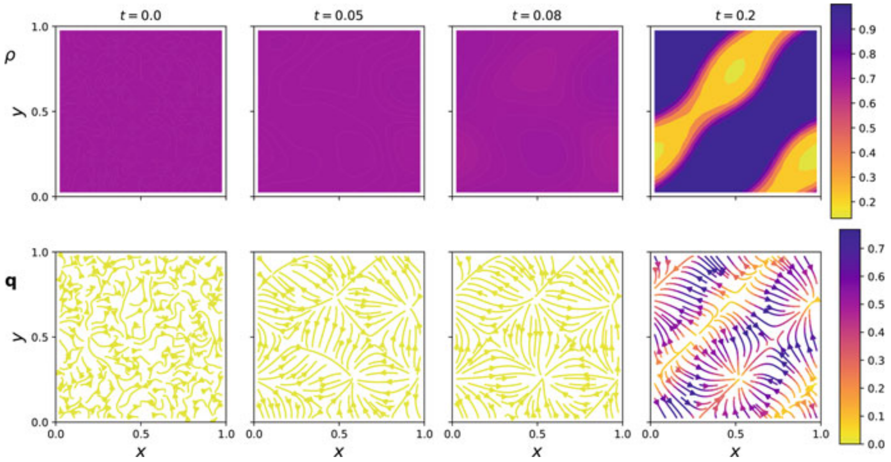


Fig. 9 Hybrid model for active particles (22) with periodic boundary conditions in $[0, 1]^2 \times [0, 2\pi]$ at different times, starting from a 3D random perturbation around the homogeneous solution. The first row shows the total density $\rho(\mathbf{x}, t)$ with mass ϕ and the second row the mean direction $(\mathbf{x}, t)$. Parameters used: $D_T = D_R = 1$, $v_0 = 60$, $\phi = 0.7$

these two regions pointing towards the dense region. We show the rescaled spatial density $\rho(\mathbf{x}, t) = \phi \int f \mathrm{d}\theta$ as well as the mean direction

$$(\mathbf{x}, t) := \frac{\int f \mathbf{e}_\theta \mathrm{d}\theta}{\int f \mathrm{d}\theta}. \tag{48}$$

Figures 10 and 11 show a comparison between the active hybrid model (22) and its corresponding passive model (30), which we obtain by setting $D_R = 1$ and a discretisation in angle with only two grid points (which define the two species r and b with respective travel directions $\theta_1 = -\pi/2$ and $\theta_2 = \pi/2$). We observe different types of segregation in each case. In the active case, we observe a blob with high density that is well-mixed in its centre (namely, orientations are uncorrelated as it corresponds to small, see bottom right plot in Fig. 10). In contrast, in the passive case, in addition to the separation into dilute and dense regions (see first row in Fig. 11), we observe a segregation of the two species within the dense phase: the red particles, which want to move downwards, are met below by a layer of blue particles, which want to move upwards (see second and third rows in Fig. 11). A similar structure in the active model (22) can be observed if the final pattern is mappable to a one-dimensional pattern, as in the case shown in Fig. 12 (which corresponds to different values of ϕ and v_0 and a different initial condition). In this case, 'left'-moving particles concentrate at one boundary and 'right'-moving particles at the other. For these same parameters, the passive model (30) displays four instead of two lanes (see Fig. 13).

Finally, we show simulation examples of the stochastic models corresponding to the active (22) and the passive (30) macroscopic models. Simulations are performed

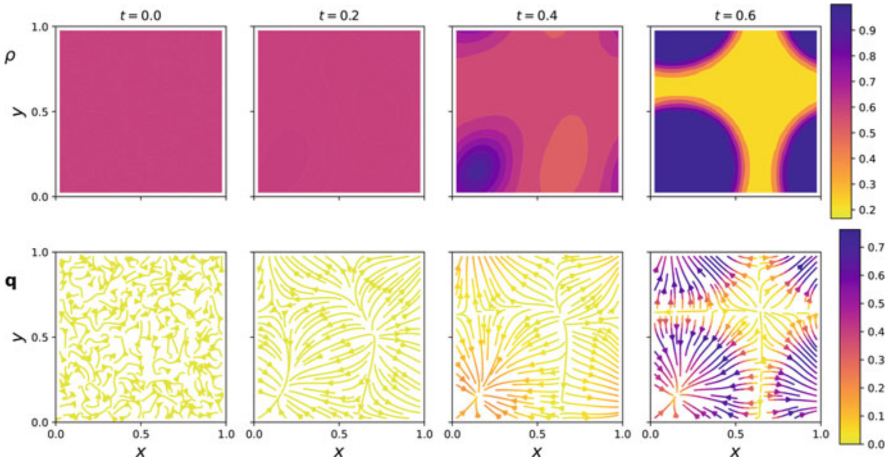


Fig. 10 Hybrid model for active particles (22) with periodic boundary conditions in $[0, 1]^2 \times [0, 2\pi]$ at different times, starting from a 3D random perturbation around the homogeneous solution. The first row shows the total density $\rho(\mathbf{x}, t)$ with mass ϕ and the second row the mean direction $\quad(\mathbf{x}, t)$ (48). Parameters used: $D_T = D_R = 1$, $v_0 = 60$, $\phi = 0.6$

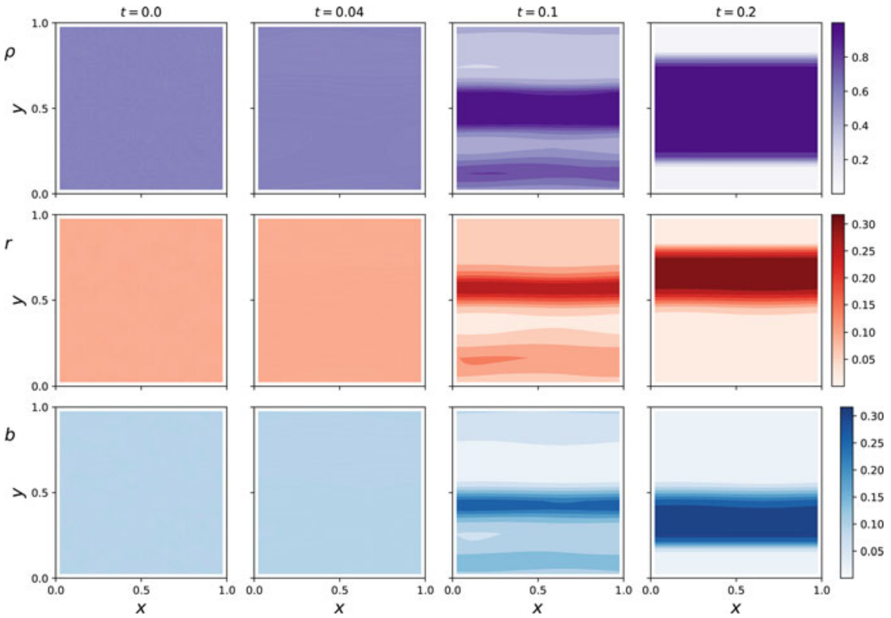


Fig. 11 Discrete model for passive particles (30) with periodic boundary conditions in $[0, 1]^2$ at different times, starting from a 3D random perturbation around the homogeneous solution. The first row shows the total density $\rho(\mathbf{x}, t) = r + b$ with mass ϕ, while the densities of the red r and the blue b species are given in the second and third rows, respectively. Parameters used: $D_T = D_R = 1$, $v_0 = 60$, $\phi = 0.6$

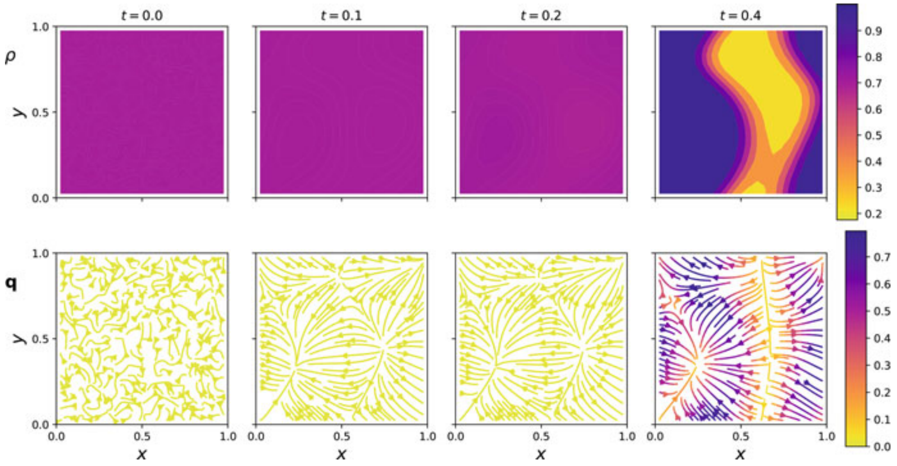


Fig. 12 Hybrid model for active particles (22) with periodic boundary conditions in $[0, 1]^2 \times [0, 2\pi]$ at different times, starting from a 3D random perturbation around the homogeneous solution. The first row shows the total density $\rho(\mathbf{x}, t)$ with mass ϕ and the second row the mean direction $(\mathbf{x}, t)$ (48). Parameters used: $D_T = D_R = 1$, $v_0 = 40$, $\phi = 0.7$

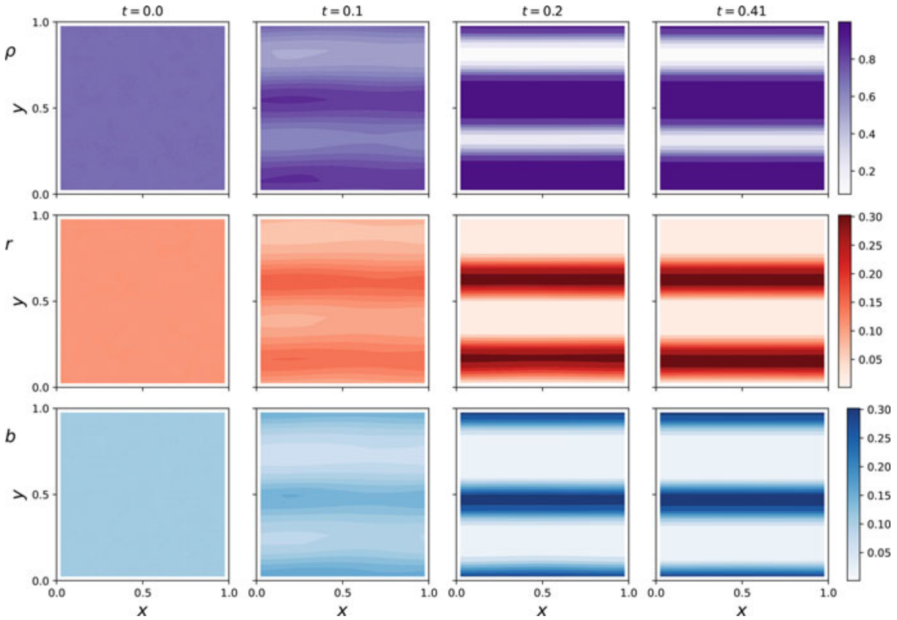


Fig. 13 Discrete model for passive particles (30) with periodic boundary conditions in $[0, 1]^2$ at different times, starting from a 2D random perturbation around the homogeneous solution. The first row shows the total density $\rho(\mathbf{x}, t) = r + b$ with mass ϕ, while the densities of the red r and the blue b species are given in the second and third rows, respectively. Parameters used: $D_T = D_R = 1$, $v_0 = 40$, $\phi = 0.7$

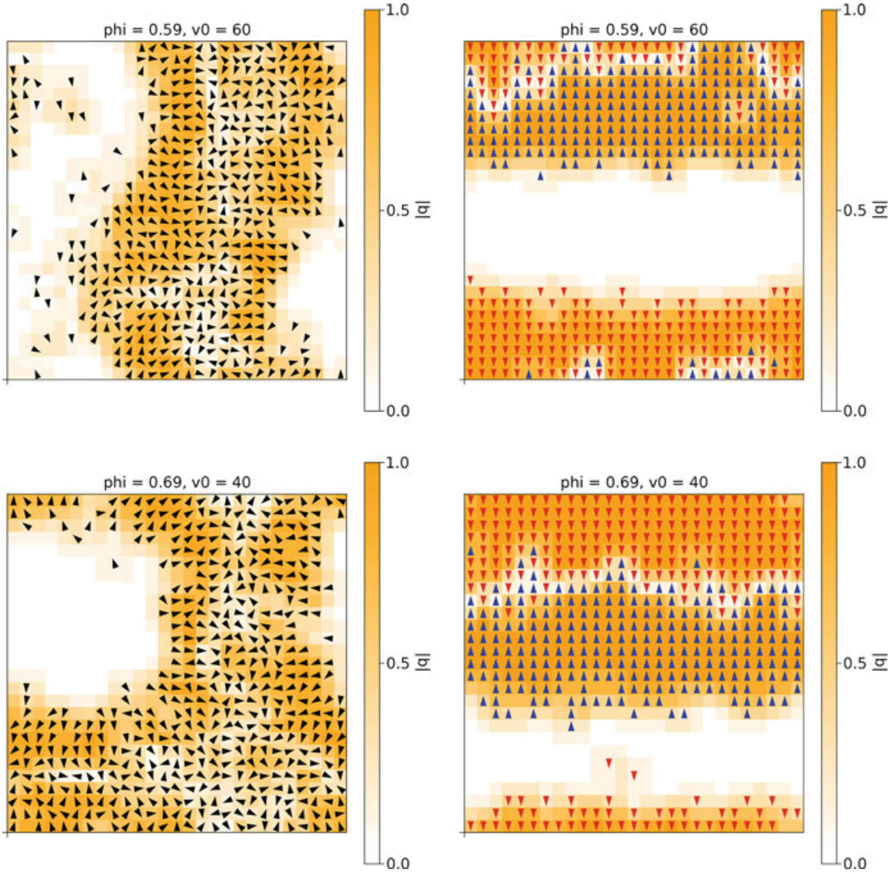


Fig. 14 Example configurations of two microscopic models corresponding to ASEP in position and different rules in orientation. (Left column): hybrid model with Brownian motion in angle with corresponding macroscopic model (22). (Right column): passive orientations as given by the initial condition, here either $\theta = \pm\pi/2$, corresponding to the macroscopic model (30) for two species of red and blue particles. Snapshots at $T = 0.5$ using fixed time steps of $\Delta t = 10^{-4}$ and parameters ϕ and v_0 as indicated above each plot. The colormap shows the strength of the polarisation $|q|$

using the agent-based modelling package Agents.jl [31] in Julia and as described in [17]. In both cases, N particles perform an ASEP [(a'), (b) and (c) mechanisms of Sect. 2.2] on a square lattice with M lattice sites such that the occupied fraction is $\phi = N/M$. In the former case, particles' orientation diffuses with D_R in $[0, 2\pi]$, so that the direction of the asymmetric jump process for each particle changes in time. In the latter case, particles are initialised as either red (pointing downwards) or blue (pointing upwards) and their orientations are fixed over time. We observe MIPS in all four cases shown in Fig. 14, with the active system displaying either a strip or blob pattern (left column) and the passive system having dilute–dense boundaries and red–blue boundaries running left to right as we had already seen in the PDE simulations (right column). The colormap in the figure shows the absolute value of the mean orientation in (48) computed using a Moore neighbourhood in each lattice: $| \quad | = 0$ in purely isotropic regions (and in empty regions) and $| \quad | = 1$ in regions with perfectly aligned particles, which happens within each segregated region of blue and red particles in the passive case and, to a lesser extent, in the boundary between dilute and dense regions in the active case.

Acknowledgments M. Bruna was partially supported by a Royal Society University Research Fellowship (grant number URF/R1/180040) and a Humboldt Research Fellowship from the Alexander von Humboldt Foundation. M. Burger acknowledges partial financial support by European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 777826 (NoMADS) and the German Science Foundation (DFG) through CRC TR 154 'Mathematical Modelling, Simulation and Optimization using the Example of Gas Networks', Subproject C06. The work of M.-T. Wolfram was partly supported by the Austrian Academy of Sciences New Frontier's grant NFG-0001. J.-F. Pietschmann thanks the DAAD for support via the PPP project 57447206.

References

1. S. Adams et al. "From a large-deviations principle to the Wasserstein gradient flow: a new micro-macro passage". In: *Communications in Mathematical Physics* 307.3 (2011), pp. 791–815.
2. L. Alasio, M. Bruna, and Y. Capdeboscq. "Stability estimates for systems with small cross-diffusion". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 52.3 (2018), pp. 1109–1135.
3. Alasio, L. et al. "Trend to equilibrium for systems with small cross-diffusion". In: *ESAIM: M2AN* 54.5 (2020), pp. 1661–1688. https://doi.org/10.1051/m2an/2020008.
4. L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
5. S.S. Andrews and D. Bray. "Stochastic simulation of chemical reactions with spatial resolution and single molecule detail". In: *Physical Biology* 1.3 (2004), pp. 137–151.
6. D. Bakry and M. Émery. "Diffusions hypercontractives". In: *Séminaire de Probabilités XIX 1983/84* Ed. by Jacques Azéma and Marc Yor. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985, pp. 177–206. ISBN: 978-3-540-39397-9.
7. C. Bechinger et al. "Active particles in complex and crowded environments". In: *Reviews of Modern Physics* 88.4 (2016), p. 045006.
8. N. Bellomo and A. Bellouquid. "On the modelling of vehicular traffic and crowds by kinetic theory of active particles". In: *Mathematical modeling of collective behaviour in socio-economic and life sciences*. Springer, 2010, pp. 273–296.
9. H. C Berg. *Random Walks in Biology*. Princeton University Press, 1993.
10. J. Bialké, H. Löwen, and T. Speck. "Microscopic theory for the phase separation of self-propelled repulsive disks". In: *EPL* 103.3 (Aug. 2013), p. 30008.
11. M. Bodnar and J. J. L. Velazquez. "Derivation of macroscopic equations for individual cell-based models: a formal approach". In: *Math. Methods Appl. Sci.* 28.15 (2005), pp. 1757–1779. ISSN: 1099-1476.
12. P C. Bressloff and B. R. Karamched. "Model of reversible vesicular transport with exclusion". In: *Journal of Physics A Mathematical General* 49.34, 345602 (2016), p. 345602. DOI: 10.1088/1751-8113/49/34/345602.

13. P. C. Bressloff and E. Levien. "Synaptic Democracy and Vesicular Transport in Axons". In: *Phys. Rev. Lett.* 114 (16 2015), p. 168101.
14. M. Bruna and S.J. Chapman. "Diffusion of multiple species with excluded-volume effects". In: *J. Chem. Phys.* 137.20 (Nov. 2012), p. 204116.
15. Maria Bruna, Martin Burger, Helene Ranetbauer, and Marie-Therese Wolfram. Asymptotic gradient flow structures of a nonlinear Fokker-Planck equation, 2017.
16. M. Bruna et al. "Cross-diffusion systems with excluded-volume effects and asymptotic gradient flow structures". In: *Journal of Nonlinear Science* 27.2 (2017), pp. 687–719.
17. M. Bruna et al. "Phase Separation in Systems of Active Particles: Modelling and Stability analysis". In: *arXiv* (2021).
18. M. Burger, Y. Dolak-Struss, and C. Schmeiser. "Asymptotic analysis of an advection-dominated chemotaxis model in multiple spatial dimensions". In: *Communications in Mathematical Sciences* 6.1 (2008), pp. 1–28.
19. M. Burger, P.A. Markowich, and J.-F. Pietschmann. "Continuous limit of a crowd motion and herding model: Analysis and numerical simulations". In: *Kinetic & Related Models* 4.4 (2011), pp. 1025–1047.
20. M. Burger and J.-F. Pietschmann. "Flow characteristics in a crowded transport model". In: *Nonlinearity* 29.11 (2016), pp. 3528–3550. DOI: 10.1088/0951-7715/29/11/3528.
21. M. Burger, B. Schlake, and M.T. Wolfram. "Nonlinear Poisson–Nernst–Planck equations for ion flux through confined geometries". In: *Nonlinearity* 25.4 (2012), p. 961.
22. M. Burger et al. "Lane Formation by Side-Stepping". In: *SIAM Journal on Mathematical Analysis* 48.2 (2016), pp. 981–1005. DOI: 10.1137/15M1033174.
23. M. Burger et al. "Lane Formation by Side-Stepping". In: *SIAM Journal on Mathematical Analysis* 48.2 (2016), pp. 981–1005.
24. M. Burger et al. "Nonlinear Cross-Diffusion with Size Exclusion". In: *SIAM J. Math. Anal.* 42.6 (2010), p. 2842.
25. J. A Carrillo et al. "Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities". In: *Monatshefte für Mathematik* 133.1 (2001), pp. 1–82.
26. José Antonio Carrillo, Yanghong Huang, and Markus Schmidtchen. Zoology of a Nonlocal Cross-Diffusion Model for Two Species. *SIAM J. Appl. Math.*, 78(2):1078 – 1104, 01 2018.
27. Michael E Cates and Julien Tailleur. Motility-Induced Phase Separation. *Annu. Rev. Cond. Ma. P.*, 6(1):219–244, 2015.
28. T Chou, K Mallick, and RKP Zia. "Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport". In: *Reports on progress in physics* 74.11 (2011), p. 116601.
29. E. N. M. Cirillo et al. "When diffusion faces drift: Consequences of exclusion processes for bi-directional pedestrian flows". In: *Physica D Nonlinear Phenomena* 413, 132651 (Dec. 2020), p. 132651.
30. E. Cristiani, B. Piccoli, and A. Tosin. *Multiscale modeling of pedestrian dynamics*. Vol. 12. Springer, 2014.
31. George Datseris, Ali R. Vahdati, and Timothy C. DuBois. Agents.jl: A performant and feature-full agent based modelling software of minimal code complexity, 2021.
32. B. Derrida. "An exactly soluble non-equilibrium system: the asymmetric simple exclusion process". In: *Physics Reports* 301.1-3 (1998), pp. 65–83.
33. B. Derrida et al. "Exact solution of a 1D asymmetric exclusion model using a matrix formulation". In: *Journal of Physics A: Mathematical and General* 26.7 (1993), pp. 1493–1517.
34. N Desai and A. M. Ardekani. "Modeling of active swimmer suspensions and their interactions with the environment". In: *Soft Matter* 13.36 (2017), pp. 6033–6050.
35. Ulrich Dobramysl, Simon Garnier, Audrey Dussutour, and Maria Bruna. Optimality of trail widths in ant foraging networks, 2021. In Preparation.
36. Audrey Dussutour, Jean-Louis Deneubourg, and Vincent Fourcassi. Amplification of individual preferences in a social context: the case of wall-following in ants. *Proceedings of the Royal Society B: Biological Sciences*, 272(1564):705–714, 2005.

37. K. Elamvazhuthi and S. Berman. “Mean-field models in swarm robotics: A survey”. In: *Bioinspiration & Biomimetics* 15.1 (2019), p. 015001.
38. J. Elgeti, R. G. Winkler, and G. Gompper. “Physics of microswimmers—single particle motion and collective behavior: a review”. In: *Reports on Progress in Physics* 78.5 (2015), p. 056601.
39. N. Gavish, P. Nyquist, and M. Peletier. “Large deviations and gradient flows for the Brownian one-dimensional hard-rod system”. In: *arXiv preprint arXiv:1909.02054* (2019).
40. Roland G Gompper G.and Winkler et al. “The 2020 motile active matter roadmap”. In: *J. Phys.: Condens. Matter* 32.19 (Feb. 2020), p. 193001.
41. D. S. Grebenkov. “Partially reflected Brownian motion: a stochastic approach to transport phenomena”. In: *Focus on probability theory* (2006), pp. 135–169.
42. D. Helbing, I. Farkas, and T. Vicsek. “Simulating dynamical features of escape panic”. In: *Nature* 407.6803 (2000), pp. 487–490.
43. D. Helbing and P. Molnar. “Social force model for pedestrian dynamics”. In: *Physical Review E* 51.5 (1995), p. 4282.
44. I. Humpert et al. “On the Role of Vesicle Transport in Neurite Growth: Modelling and Experiments”. In: *arXiv:1908.02055 [q-bio]* (2019).
45. A. Jüngel. “The boundedness-by-entropy method for cross-diffusion systems”. In: *Nonlinearity* 28.6 (2015), p. 1963.
46. Y.-E. Keta et al. “Collective motion in large deviations of active particles”. In: *Phys. Rev. E* 103.2 (2021), p. 022603.
47. A. Kirchner and A. Schadschneider. “Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics”. In: *Physica A: Statistical Mechanics and its Applications* 312.1 (2002), pp. 260–276. issn: 0378-4371.
48. M. Kourbane-Houssene et al. “Exact Hydrodynamic Description of Active Lattice Gases”. In: *Phys. Rev. Lett.* 120 (26 2018), p. 268003.
49. P. L. Lions and A. S. Sznitman. “Stochastic differential equations with reflecting boundary conditions”. In: *Communications on Pure and Applied Mathematics* 37.4 (1984), pp. 511–537.
50. J. Maas. “Gradient flows of the entropy for finite Markov chains”. In: *Journal of Functional Analysis* 261.8 (2011), pp. 2250–2292.
51. B. Maury and S. Faure. *Crowds in Equations: An Introduction to the Microscopic Modeling of Crowds*. World Scientific, 2018.
52. Bertrand Maury et al. “Handling congestion in crowd motion modeling”. In: *Networks and Heterogeneous Media* 6.3 (2011), p. 485.
53. L.G. Nava, R. Großmann, and F. Peruani. “Markovian robots: Minimal navigation strategies for active particles”. In: *Physical Review E* 97.4 (2018), p. 042604.
54. B. J. Nelson, I. K. Kaliakatsos, and J. J. Abbott. “Microrobots for minimally invasive medicine”. In: *Annual review of biomedical engineering* 12 (2010), pp. 55–85.
55. M.A. Peletier et al. “Jump processes as Generalized Gradient Flows”. In: (2020). arXiv: `2006.10624 [math.AP]`.
56. J.-F. Pietschmann and B. Schlake. “Lane formation in a microscopic model and the corresponding partial differential equation”. In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE. 2011, pp. 173–179.
57. A. Pilipenko. *An introduction to stochastic differential equations with reflection*. Vol. 1. Universitätsverlag Potsdam, 2014.
58. T. Ralph, S. W Taylor, and M. Bruna. “One-dimensional model for chemotaxis with hard-core interactions”. In: *Phys. Rev. E* 101.2 (2020), p. 022419.
59. H. Rost. “Diffusion de sphéres dures dans la droite réelle: comportement macroscopique et équilibre local”. In: *Séminaire de Probabilités XVIII 1982/83* Springer, 1984, pp. 127–143.
60. F. Schmidt et al. “Light-controlled assembly of active colloidal molecules”. In: *The Journal of chemical physics* 150.9 (2019), p. 094905.
61. Markus Schmidtchen, Maria Bruna, and S Jonathan Chapman. Excluded volume and order in systems of Brownian hard needles, 2021. In Preparation.
62. J. Schöberl. “NETGEN An advancing front 2D/3D-mesh generator based on abstract rules”. In: *Computing and Visualization in Science* 1.1 (1997), pp. 41–52.

63. M. J. Simpson, K. A. Landman, and B. D Hughes. “Multi-species simple exclusion processes”. In: *Physica A* (2009).
64. A. V. Skorokhod. “Stochastic Equations for Diffusion Processes in a Bounded Region”. In: *Theory of Probability & Its Applications* 6.3 (1961), pp. 264–274. DOI: 10.1137/1106035.
65. T. Speck et al. “Dynamical mean-field theory and weakly non-linear analysis for the phase separation of active Brownian particles.” In: *J. Chem. Phys.* 142.22 (June 2015), p. 224109.
66. J. Tailleur and M. E. Cates. “Statistical Mechanics of Interacting Run-and-Tumble Bacteria”. In: *Phys. Rev. Lett.* 100 (21), p. 218103. DOI: 10.1103/PhysRevLett.100.218103.
67. D. B. Wilson, H. Byrne, and M. Bruna. “Reactions, diffusion, and volume exclusion in a conserved system of interacting particles”. In: *Phys. Rev. E* 97.6 (June 2018), p. 062137.
68. J.A. Wood. “A totally asymmetric exclusion process with stochastically mediated entrance and exit”. In: *Journal of Physics A: Mathematical and Theoretical* 42.44 (2009), p. 445002.
69. A. Zöttl and H. Stark. “Emergent behavior in active colloids”. In: *Journal of Physics: Condensed Matter* 28.25 (2016), p. 253001.

Mathematical Modeling of Cell Collective Motion Triggered by Self-Generated Gradients

Vincent Calvez, Mete Demircigil, and Roxana Sublet

Abstract Self-generated gradients have attracted a lot of attention in the recent biological literature. It is considered as a robust strategy for a group of cells to find its way during a long journey. This note is intended to discuss various scenarios for modeling traveling waves of cells that constantly deplete a chemical cue and so create their own signaling gradient all along the way. We begin with one famous model by Keller and Segel for bacterial chemotaxis. We present the model and the construction of the traveling wave solutions. We also discuss the limitation of this approach and review some subsequent work addressing stability issues. Next, we review two relevant extensions, which are supported by biological experiments. They both admit traveling wave solutions with an explicit value for the wave speed. We conclude by discussing some open problems and perspectives, and particularly a striking mechanism of speed determinacy occurring at the back of the wave. All the results presented in this note are illustrated by numerical simulations.

1 Introduction

It has been now 50 years that Evelyn F. Keller and Lee A. Segel published their article "*Traveling bands of chemotactic bacteria: A theoretical analysis*" [40], which is part of a series of works about the modeling of chemotaxis in bacteria *Esherichia coli* and amoebae *Dictyostelium discoideum* (shortnamed as Dicty in the following) [37, 38, 39, 40]. This article described in a simple and elegant way the propagation of chemotactic waves of *E. coli* in a one-dimensional space, echoing the remarkable experiments by Adler performed in a capillary tube [1].

V. Calvez (✉) · M. Demircigil
Institut Camille Jordan (ICJ), UMR 5208 CNRS & Université Claude Bernard Lyon 1, and Equipe-Projet Inria Dracula, Lyon, France
e-mail: vincent.calvez@math.cnrs.fr; mete.demircigil@univ-lyon1.fr

R. Sublet
Institut Camille Jordan (ICJ), UMR 5208 CNRS & Université Claude Bernard Lyon 1, Lyon, France

N. Bellomo et al. (eds.), *Active Particles, Volume 3*, Modeling and Simulation in Science, Engineering and Technology, https://doi.org/10.1007/978-3-030-93302-9_3

In the present contribution, the seminal ideas of Keller and Segel are discussed from a modern perspective, after half a century of intense activity at the interface of mathematics and biology. Our goal is not to review exhaustively various directions of research in the modeling of chemotaxis. Our narrow objective consists in setting the focus on the notion of *self-generated gradients* (SGG), which has recently shed a new light on several biological processes, both in bacterial collective motion and in some aspects of developmental biology [69, 67]. SGG are at the heart of the model in [40], in which cells create their own signaling gradient by consuming some nutrient while moving collectively from one side of the domain to the other. There, collective motion results from the averaged biases in the individual trajectories, in response to nutrient heterogeneities, a process called chemotaxis. This concept of SGG can be generalized to any situation where the signal depletion and chemotaxis functions overlap within the same cells [58, 68, 69].

SGG in Waves of Bacteria The work of Keller and Segel has initiated a wealth of studies on bacterial chemotaxis. We refer to the comprehensive review of Tindall et al. [65], and also the recent studies [29, 17] for new biological questions in this topic. Most of the works discussed in this note consider short time experiments, or experiments at low level of nutrients, neglecting the effect of cell division. This makes a clear distinction between SGG and reaction–diffusion waves, as the celebrated Fisher/Kolmogorov–Petrovsky–Piskunov (F/KPP) equation [27, 41, 3]. For this reason, we shall not comment further about the numerous modeling contributions following the patterns reported by Budrene and Berg [8, 9, 7] (ring expansion followed by formation of bacteria spots with remarkable symmetries). Chemotaxis has been shown to be crucial in the emergence of such patterns. However, the dynamics of ring expansion are mainly driven by growth and diffusion such as described by F/KPP (but see [17] for a recent study where chemotaxis has been shown to enhance range expansion).

There exist many modeling contributions of chemotaxis in bacteria [65, 31], with a particular emphasis on the derivation of macroscopic models from individual rules through kinetic transport equations; see, e.g., [2, 49, 50, 15, 26, 14, 4]. In contrast, the number of contributions about mathematical analysis of traveling waves without growth beyond [40] is relatively scarce. We refer to [33], for an (algebraic) extension of [40] with more general chemotaxis functions and uptake rates. We also refer to the series of articles by Z.A. Wang and co-authors; see [73] for a preliminary review and below for further discussion.

SGG in Development and Cancer In developmental biology, cell movement over long distances is mediated by navigating cues, including chemotactic factors [45]. It is commonly postulated that external, pre-patterned gradients, drive cellular migration. One of the key conceptual advantages of SGG is to free the developmental process from the requirement of pre-imposed long-range chemoattractant gradients. In contrast, SGG travel together with the cells, so that they can experience similar environmental conditions (chemical concentration, gradient steepness) all over the journey. This is thought to provide robustness to the developmental system [68, 67].

Recently, SGG have been shown to occur during embryogenesis, and in particular during the initiation of the posterior lateral line in zebrafish [22, 72]. More precisely, migrating cell cohorts (consisting of approximately a hundred of cells) can generate and sustain gradients of chemoattractants across their length. This experimental work is of great importance as being the first proof of the occurrence of SGG in vivo.

Self-generated gradients are also under investigation during cancer invasion and metastasis. This includes modeling in silico (see [59] and references therein), and experiments with cell cultures in vitro [47]. In particular, we highlight the work of [58], in which an astonishing self-guidance strategy in cancer epithelial cell populations was unravelled. In fact, cells were put in microfluidic mazes, without any pre-existing external gradients. Most of them could find their way out of the mazes by generating their own navigating cues. Experimental studies with increasingly complex mazes were also performed with Dicty cells, with quite remarkable outcomes [70].

Plan and Purpose of the Chapter In Sect. 2.1 we recall the basic construction of traveling waves in the seminal article [40]. The lack of positivity of the chemical concentration is illustrated by some numerical simulations (Sect. 2.2). The issue of instability is also reviewed. Section 2.3 briefly presents some possible variations of the original article from the literature. It is one of the main goals of the present contribution to discuss in detail two possible extensions that are biologically relevant (that is, supported by experiments). Section 3 contains an overview of past work where another attractant signal is added to prevent cell dispersion during propagation. This results in competing cell fluxes, with stronger advection at the back of the wave than at the edge. Section 4 reports on a piece of recent work including signal-dependent phenotypical switch (division/migration). This results in a wave sustained by cell division restricted to the edge.

All mathematical results proven here are simple, namely involving explicit construction of one-dimensional traveling waves (whose respective stabilities are supported by numerical simulations of the Cauchy problems). The last construction is original, up to our knowledge; see Theorem 4.2. It could be of interest for experts in reaction–diffusion equations, as it exhibits a possibly new phenomenon of selection of the minimal speed at the back of the wave.

2 The Keller–Segel Model and Variations

2.1 The Construction of Waves by Keller and Segel

In this section, we recall briefly the model and analysis in [40]. The cell density (bacteria) is denoted by $\rho(t, x)$, for time $t > 0$, and position along the channel axis $x \in \mathbb{R}$, whereas the concentration of the signaling molecule is denoted by $S(t, x)$.

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}\left(-d\dfrac{\partial \rho}{\partial x} + \chi \rho \dfrac{\partial \log S}{\partial x}\right) = 0\,, \\ \dfrac{\partial S}{\partial t} = D\dfrac{\partial^2 S}{\partial x^2} - k\rho\,. \end{cases} \tag{2.1}$$

The equation on ρ combines unbiased (diffusive) motion with directed motion in response to the logarithmic signaling gradient (see below for further discussion about this specific choice), with intensity $\chi > 0$.

On the one hand, the equation on ρ is conservative, and the total mass of cells in the channel, which is an invariant of the system, is denoted by M, so that $M = \int_{\mathbb{R}} \rho(0, z)\, dz = \int_{\mathbb{R}} \rho(t, z)\, dz$. On the other hand, the chemical concentration decays globally in time, and the limiting value at ∞ is denoted by S_{init}, which can be viewed as the initial, homogeneous, concentration in the channel associated with the Cauchy problem.

Noticeably, the consumption term in the equation on S, namely $-k\rho$, does not involve S itself, precluding any guarantee about the positivity of S in the long time. Nevertheless, the existence of positive traveling wave solutions $\rho(x-ct)$, $S(x-ct)$ was established in [40] by means of explicit computations, in the absence of signal diffusion $D = 0$ (for mathematical purposes), and with the condition $\chi > d$. The wave under interest has the following structure: $\rho \in L^1_+(\mathbb{R})$, with $\lim_{z\to\pm\infty} \rho(z) = 0$, and $S \in L^\infty_+(\mathbb{R})$ is increasing with $\lim_{z\to-\infty} S(z) = 0$, and $\lim_{z\to+\infty} S(z) = S_{\text{init}}$, the reference value of the chemical concentration.

Theorem 2.1 (Keller and Segel [40]) *Assume $D = 0$, and $\chi > d$. Then, there exist a speed $c > 0$ (depending on M, k, and S_{init}, but not on χ nor on d) and a stationary solution of* (2.1) *in the moving frame $(\rho(x-ct), S(x-ct))$, such that ρ is positive and integrable, $\int_{\mathbb{R}} \rho(z)\, dz = M$, and S is increasing between the following limiting values:*

$$\begin{cases} \lim_{z\to-\infty} S(z) = 0\,, \\ \lim_{z\to+\infty} S(z) = S_{\text{init}}\,. \end{cases}$$

Before we recall briefly the construction of the wave solution, let us comment on the value of the wave speed c that can be directly obtained from the second equation in (2.4), whatever the value of $D \geq 0$ is. Indeed, the equation in the moving frame reads

$$-c\frac{dS}{dz} = D\frac{d^2 S}{dz^2} - k\rho\,.$$

By integrating this equation over the line, and using the extremal conditions at $\pm\infty$ (that can be verified a posteriori), we find

$$cS_{\text{init}} = k\int_{\mathbb{R}} \rho(z)\, dz = kM\,. \tag{2.2}$$

Strikingly, the wave speed depends only on the dynamics of establishment of the gradient. In particular, it does not depend on the intensity of the chemotactic response χ. This is in contrast with several conclusions to be drawn from alternative models in the sequel (see Sects. 3 and 4).

Proof The speed c is given a priori by the relationship (2.2).

The first step in the construction of traveling wave solutions is the zero-flux condition in the moving frame $z = x - ct$, namely

$$-c\rho - d\frac{d\rho}{dz} + \chi\rho\frac{d\log S}{dz} = 0 \quad \Leftrightarrow \quad \frac{d\log\rho}{dz} = -\frac{c}{d} + \frac{\chi}{d}\frac{d\log S}{dz}$$
$$\Leftrightarrow \quad \rho(z) = a\exp\left(-\frac{c}{d}z + \frac{\chi}{d}\log S\right)\,,$$

where a is a (positive) constant of integration. The second step consists in solving the following ODE (assuming $D = 0$):

$$c\frac{dS}{dz} = ka\exp\left(-\frac{c}{d}z + \frac{\chi}{d}\log S\right)$$
$$\Leftrightarrow \left(1 - \frac{\chi}{d}\right)^{-1}\left(S_{\text{init}}^{1-\frac{\chi}{d}} - S(z)^{1-\frac{\chi}{d}}\right) = \frac{kad}{c^2}\exp\left(-\frac{c}{d}z\right)\,.$$

By re-arranging the terms, we obtain

$$\left(\frac{S(z)}{S_{\text{init}}}\right)^{1-\frac{\chi}{d}} = 1 + \left(\frac{\chi}{d} - 1\right)\left(\frac{kad}{c^2}S_{\text{init}}^{\frac{\chi}{d}-1}\right)\exp\left(-\frac{c}{d}z\right)\,.$$

Suppose that $\chi < d$, then the right-hand side goes to $-\infty$ as $z \to -\infty$, which is a contradiction. Hence, the calculations make sense only if $\chi > d$. By translational invariance, the constant a can be chosen so as to cancel the prefactor in the right-hand side (provided $\chi > d$), yielding the simple expression:

$$\frac{S(z)}{S_{\text{init}}} = \left(1 + \exp\left(-\frac{c}{d}z\right)\right)^{\frac{d}{d-\chi}}\,. \tag{2.3}$$

The corresponding density profile is

$$\rho(z) = a'\exp\left(-\frac{c}{d}z\right)\left(1 + \exp\left(-\frac{c}{d}z\right)\right)^{\frac{\chi}{d-\chi}}\,, \tag{2.4}$$

for some constant a', that can be determined explicitly through the conservation of mass. □

2.2 Positivity and Stability Issues

Despite its elegance, the previous construction suffers from two drawbacks. First of all, the positivity of the signal concentration S is not guaranteed in the Cauchy problem. Actually, numerical solutions soon break down because of this positivity issue. This occurs starting from a generic initial data (Fig. 1a), and even from the traveling wave solution $(\rho(z), S(z))$ given by the expressions (2.3)–(2.4), after accumulation of numerical errors (Fig. 1b). Nevertheless, the positivity can be manually rescued by setting $S_{n+1} = \max(S_n, \epsilon)$ for some arbitrary threshold $\epsilon \ll 1$, as suggested in [32]. In that case, the wave seems to propagate in a stable way in the long term; see Fig. 1c.

Second, and somewhat related, is the problem of stability of the wave constructed in Sect. 2.1. Linear stability was addressed first in [53], where it was proven that the spectral problem admits no *real* positive eigenvalue. However, the linearized problem is not self-adjoint, so that this preliminary result is largely incomplete from the perspective of stability. Few years later, it was proven in [48] that the (essential) spectrum of the linear operator intersects the right half-plane, meaning that the wave is linearly unstable. The authors proved a refined instability result, when perturbations are restricted to a class of exponentially decreasing functions. Noticeably, their results cover both $D = 0$ and $D > 0$. This analysis has been largely extended in [18, 19] where it was proven that the wave is either transiently (convectively) unstable; that is, the spectrum is shifted in the open left half plane in a two-sided exponentially weighted function space [55], when $\chi > d$ is not too large, but it is absolutely unstable when χ is above some threshold, that is, $\frac{\chi}{d} > \beta^0_{\text{crit}}(D)$, where, e.g., $\beta^0_{\text{crit}}(0)$ is the unique real root above one of an explicit 10th order polynomial; see [19, Theorem 2.1].

Recently, it has been established the existence and nonlinear stability of stationary solutions for the problem (2.1) set on a half-line $\{x > 0\}$, with, respectively, Neumann boundary condition for ρ, and positive Dirichlet boundary condition for S at the origin [13]. The motivation comes from the study of spike solutions stabilized by a sustained amount of chemical concentration at the boundary. The stability result in [13] imposes quite stringent conditions on the decay of the initial data at $+\infty$. Nevertheless, local stability of the stationary spike does not preclude loss of positivity in the numerics when initiating the Cauchy problem with initial conditions far from equilibrium; see Fig. 2.

Remark 2.2 Many of the references mentioned above also discuss and analyze the case of a degenerate consumption rate $\frac{\partial S}{\partial t} = D\frac{\partial^2 S}{\partial x^2} - k\rho S^m$ $(m < 1)$, without changing much of the global picture.

The case $m = 1$ differs significantly, however. It can be viewed directly from the case $D = 0$ that the logarithmic gradient of the putative wave in the moving frame, that is, $\frac{d \log S}{dz}$ cannot have a positive limit as $z \to -\infty$, simply because it satisfies the relationship $-c\frac{d \log S}{dz} = -k\rho$, the latter being integrable. Consequently, advection cannot balance diffusion at $-\infty$, preventing the existence of a traveling

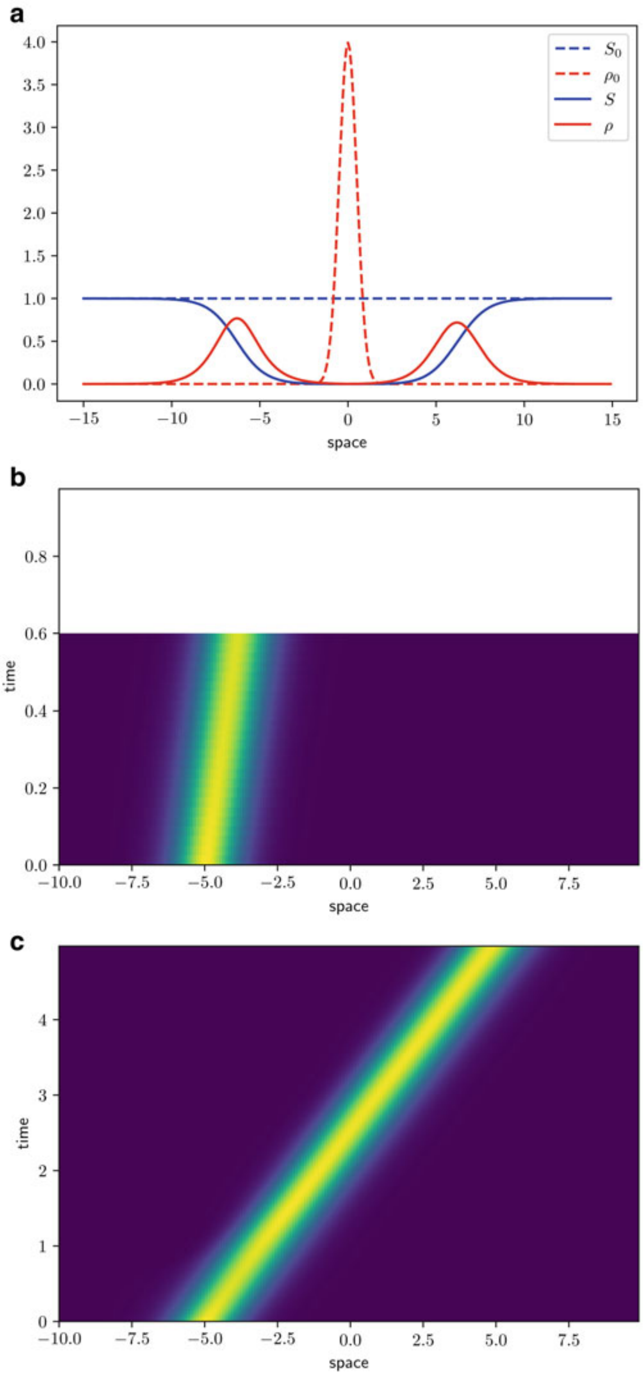


Fig. 1 Positivity and stability issues in the numerical simulations of (2.1). (**a**) Starting from a generic initial data, the numerical scheme quickly breaks down because the signal becomes negative at some point. The initial condition is shown in dashed line, and the final state in plain line (last time before numerical breakdown). (**b**) Aligning the initial data on the exact density and signal profiles $(\rho(z), S(z))$, (2.3)–(2.4), yields the same conclusion. The cell density is shown in space/time. The numerical breakdown occurs at approximately $t = 0.6$. (**c**) The propagation of the wave can be rescued by setting manually $S_{n+1} = \max(S_n, 1E - 12)$ after each time step, as in [32]. For all the figures, the parameters are $(d = 1, \chi = 2, D = 0, k = 1)$

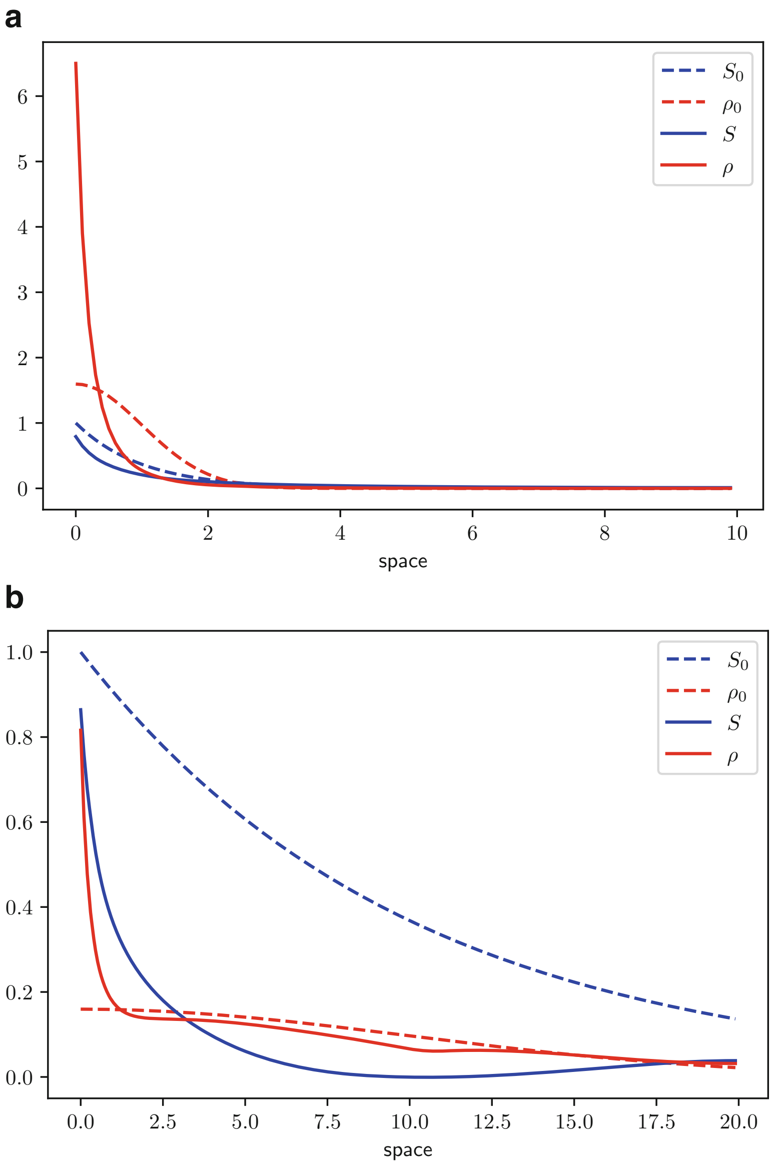


Fig. 2 Numerical solutions of (2.1) with, respectively, Neumann boundary condition for ρ, and positive Dirichlet boundary condition for S at the origin. (**a**) Local stability, as established in [13] is illustrated numerically, for an initial condition chosen near the stationary state, and a relatively large diffusion of the chemical ($d = 1, \chi = 2, D = 1, k = 1$). (**b**) Nonetheless, the numerical solution may become nonpositive when the initial condition is far from the stationary state, and diffusion of the chemical is not too large ($d = 1, \chi = 2, D = 0.25, k = 1$). For each figure, the initial condition is shown in dashed line, and the final state in plain line (last time before numerical breakdown in (**b**))

wave. The same conclusion holds in the case $D > 0$, for which $u = \frac{d \log S}{dz}$ is a homoclinic orbit of the following first-order equation:

$$\frac{du}{dz} = -\frac{c}{D}u - u^2 + \frac{k}{D}\rho \,,$$

that leaves the origin $u = 0$ at $z = -\infty$ and gets back to the origin $u = 0$ at $z = +\infty$; see [10, Proposition 6.3].

2.3 Variations on the Keller–Segel Model

As mentioned above, the seminal work [40] gave rise to a wealth of modeling and analysis of traveling bands of bacteria. Many extensions were proposed soon after Keller and Segel's original paper, with various sensitivity functions (other than the logarithmic sensitivity), and various consumption rates. The models have the following general form:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}\left(-d\dfrac{\partial \rho}{\partial x} + \rho \boldsymbol{\chi}\left(S, \dfrac{\partial S}{\partial x}\right)\right) = 0 \,, \\ \dfrac{\partial S}{\partial t} = D\dfrac{\partial^2 S}{\partial x^2} - \mathbf{k}(S, \rho) \,, \end{cases} \tag{2.5}$$

where the chemotactic sensitivity $\boldsymbol{\chi}$ can be a function of both the signal concentration and its gradient (as well as the diffusion coefficient d—dependency not reported here for the sake of clarity). These variations were nicely reviewed by Tindall et al. [65], and we are not going to comment them, but the contribution of Rivero et al. [51]. The latter follows the approach of Stroock [64] and Alt [2]. These approaches make the connection between the individual response of bacteria to space–time environmental heterogeneities and the macroscopic flux, hence making sense of the aforementioned averaging, by means of individual biases in the trajectories (see, e.g., [49, 50, 15, 26, 14, 4], and more specifically [24, 56, 60, 74] for bacterial populations). Interestingly, Rivero et al. postulate a chemotactic advection speed $\boldsymbol{\chi}$, which is nonlinear with respect to the chemical gradient at the macroscopic scale, namely

$$\boldsymbol{\chi}\left(S, \frac{\partial S}{\partial x}\right) = \chi \tanh\left(f(S)\frac{\partial S}{\partial x}\right) \,, \tag{2.6}$$

where f is a decreasing function containing the details of signal integration by a single cell.

Up to our knowledge, none of the models in the long list of existing variations could exhibit traveling waves while preserving positivity of S and keeping the total mass $\int_{\mathbb{R}} \rho$ constant (that is, ignoring growth). The minimal requirement for ensuring positivity would essentially be that the uptake function $\mathbf{k}(S, \rho)$ is dominated by S at small concentration, typically: $\limsup_{S\to 0} \frac{\mathbf{k}(S,\rho)}{S} < \infty$. However, this intuitively leads to a shallow (logarithmic) gradient at the back of the wave, unable to guarantee the effective migration of cells left behind; see Remark 2.2. Cell leakage has long been identified in the biological literature but not considered as a major issue; see, for instance, a discussion in [29], and also the addition of a linear growth term in [57] so that the loss of cells at the back is qualitatively compensated by cell division (for a realistic value of the division rate).

It is interesting to discuss the natural choice $-\mathbf{k}(S, \rho) = -kS\rho$ (combined with logarithmic sensitivity), which has been widely studied using tools from hyperbolic equations (after performing the Hopf–Cole transformation) by Z.A. Wang and co-authors; see the review [73], and further stability results in [35, 42]. The issue of shallow gradients is overcome by the boundary conditions at infinity, ρ being uniformly positive at least on one side. Clearly, the traveling wave solutions are not integrable. This hints to the conflict of conservation of mass and chemical positivity, which seem not concilable.

This leakage effect is a major mathematical issue, because most of the analytical studies build upon the existence of a wave speed and a wave profile, which is stationary in the moving frame.

2.4 *Beyond the Keller–Segel Model: Two Scenarios for SGG*

In the next two sections, we discuss two relevant modeling extensions, motivated by biological experiments, for which traveling waves exist and are expected to be stable. In the first scenario, cell leakage is circumvented by enhanced advection at the back of the wave, with an asymptotic constant value of the transport speed at $-\infty$. In the second scenario, cell leakage occurs, but it is naturally compensated by growth at the edge of the propagating front.

For each scenario, we discuss briefly the biological motivations. Then we present the explicit construction of the traveling wave solutions, together with the formula for the wave speed. When possible, we discuss the connections with some other works in the literature.

3 Scenario 1: Strongest Advection at the Back

In this section, we present some study performed a decade ago, revisiting original Adler's experiment; see Fig. 3. Inspired by massive tracking analysis, Saragosti et al. [56] proposed a simple model for the propagation of chemotactic waves of bacteria, including two signals (see also [75] for an analogous approach developed independently at the same time). The macroscopic model is the following:

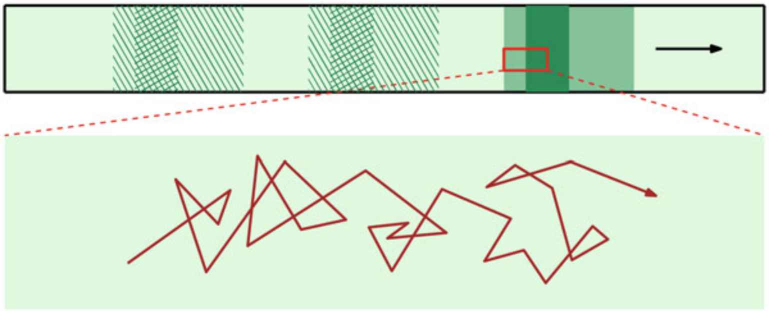


Fig. 3 Cartoon of the experiments performed in [56] and [57]. A band of bacteria is traveling from left to right in a microfluidic channel. Videomicroscopy allows tracking individual trajectories inside the wave, revealing heterogeneous behaviors: biases are stronger at the back of the wave than at the edge

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}\left(-d\dfrac{\partial \rho}{\partial x} + \rho\left(\chi_S \mathrm{sign}\left(\dfrac{\partial S}{\partial x}\right) + \chi_A \mathrm{sign}\left(\dfrac{\partial A}{\partial x}\right)\right)\right) = 0\,, \\ \dfrac{\partial S}{\partial t} = D_S \dfrac{\partial^2 S}{\partial x^2} - \mathbf{k}(S,\rho)\,, \\ \dfrac{\partial A}{\partial t} = D_A \dfrac{\partial^2 A}{\partial x^2} + \beta\rho - \alpha A\,. \end{cases} \tag{3.1}$$

As compared to (2.5), it is supplemented with a second chemical signal, A, which plays the role of a communication signal released by the cell population (hence, the source term $+\beta\rho$) and naturally degraded at a constant rate $\alpha > 0$. Indeed, bacteria are known to secrete amino acids, which play the role of a chemo-attractant as part of a positive feedback loop [5, 46].

Moreover, bacteria are assumed to respond to the signal in a binary way at the macroscopic scale: the advection speed associated with each signal (S, A) can take only two values, respectively, $\pm\chi_S$ and $\pm\chi_A$, depending on the direction of the gradients. Then, the total advection speed is simply the sum of the two components. This was derived in [56] from a kinetic model at the mesoscopic scale, assuming a strong amplification during signal integration; see also [10] for a discussion. This can be viewed as an extremal choice of the advection speed proposed by Rivero et al. [51], in the regime $f \to +\infty$ (2.6). The biophysical knowledge about the details of signal integration in bacteria *E. coli* have increased in the meantime [66, 36, 34, 61]. Actually, the logarithmic sensing is a good approximation in a fairly large range of signal concentrations. However, we retain this simple, binary, choice for theoretical purposes.

As for the Keller–Segel model, traveling waves for (3.1) have the great advantage of being analytically solvable, essentially because the problem reduces to an equation with piecewise constant coefficients. Introduce again the variable $z = x - ct$ in the moving frame at (unknown) speed c. Then, we have the following result:

Theorem 3.1 (Saragosti et al. [56]) *There exist a speed $c > 0$ and a positive limit value $S_- < S_{\mathrm{init}}$, such that the system (3.1) admits a stationary solution in the moving frame $(\rho(x-ct), S(x-ct), A(x-ct))$, such that ρ is positive and integrable, $\int_{\mathbb{R}} \rho(z)\, dz = M$, A decays to zero on both sides, and S is increasing between the following limiting values:*

$$\begin{cases} \lim_{z\to-\infty} S(z) = S_- \,, \\ \lim_{z\to+\infty} S(z) = S_{\mathrm{init}} \,. \end{cases}$$

Moreover, the speed $c > 0$ is determined by the following implicit relation:

$$\chi_S - c = \chi_A \frac{c}{\sqrt{c^2 + 4\alpha D_A}} \,. \tag{3.2}$$

Proof Contrary to the proof of Theorem 2.1, the wave speed c cannot be computed by a direct argument.

As a preliminary step, we should prescribe the environmental conditions, as they are expected heuristically to be seen by the bacteria; see Fig. 4. On the one hand, we seek an increasing profile S, hence sign $\left(\frac{dS}{dz}\right) = +1$, and the equation on the density profile ρ is decoupled from the dynamics of S. On the other hand, we assume that the communication signal A reaches a unique maximum, which can be set at $z = 0$ by translational invariance. The validation of this ansatz, a posteriori, will set the equation for c (3.2).

The equation for ρ has now piecewise constant coefficients in the moving frame:

$$-c\frac{d\rho}{dz} + \frac{d}{dz}\left(-d\frac{d\rho}{dz} + \rho\left(\chi_S + \chi_A \mathrm{sign}\ (-z)\right)\right) = 0\,.$$

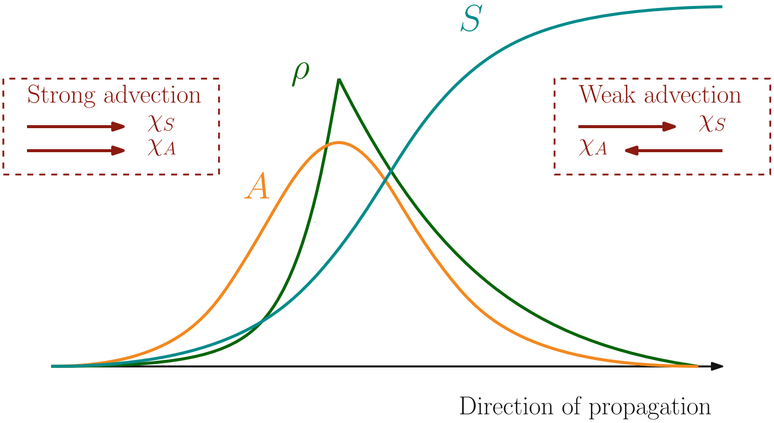


Fig. 4 Sketch of the chemical environment viewed by the cell density in model (3.1). It is characterized by stronger advection at the back (the two signals have the same orientation), than at the edge (the two gradients have opposite orientations). When chemotactic speeds coincide ($\chi_S = \chi_A$), then we simply have diffusion on the right side of the peak

Hence, ρ is a combination of two exponential functions:

$$\rho(z) = \begin{cases} \exp(\lambda_- z) & \text{for } z < 0, \quad \lambda_- = \dfrac{-c + \chi_S + \chi_A}{d} \quad \text{(signals are aligned)}, \\ \exp(-\lambda_+ z) & \text{for } z > 0, \quad \lambda_+ = \dfrac{c - \chi_S + \chi_A}{d} \quad \text{(signals are competing)}. \end{cases}$$

Next, the attractant concentration A can be computed explicitly, by convolving the source term $\beta\rho$ with the fundamental solution of the elliptic operator $-c\frac{d}{dz} - D_A\frac{d^2}{dz^2} + \alpha$, denoted by $\mathcal{A}$, that is, $A = \beta\mathcal{A} * \rho$. Coincidentally, $\mathcal{A}$ shares the same structure as ρ, namely $\mathcal{A}(z) = a_0 \exp(\mu_- z)$ for $z < 0$ and $\mathcal{A}(z) = a_0 \exp(-\mu_+ z)$ for $z > 0$, with $\mu_\pm = \frac{1}{2D_A}\left(\pm c + \sqrt{c^2 + 4\alpha D_A}\right)$, and a_0 is a normalizing factor.

It remains to check the preliminary ansatz, that is, A changes monotonicity at $z = 0$. A straightforward computation yields

$$\frac{dA}{dz}(0) = \beta a_0 \left(-\frac{1}{1 + \lambda_-/\mu_+} + \frac{1}{1 + \lambda_+/\mu_-} \right).$$

Therefore, the construction is complete, provided $\lambda_-\mu_- = \lambda_+\mu_+$, which is equivalent to (3.2). □

To partially conclude, let us highlight the fact that cohesion in the wave is guaranteed by the local aggregation signal A. To put things the other way around, in the absence of the driving signal S, the cells can aggregate thanks to the secretion of A, and the density reaches a stationary state (standing wave). In turn, this cohesive state can travel (with some deformation) in the presence of the (self-generated) driving signal S, see Fig. 5 for a numerical illustration. To make the link with SGG

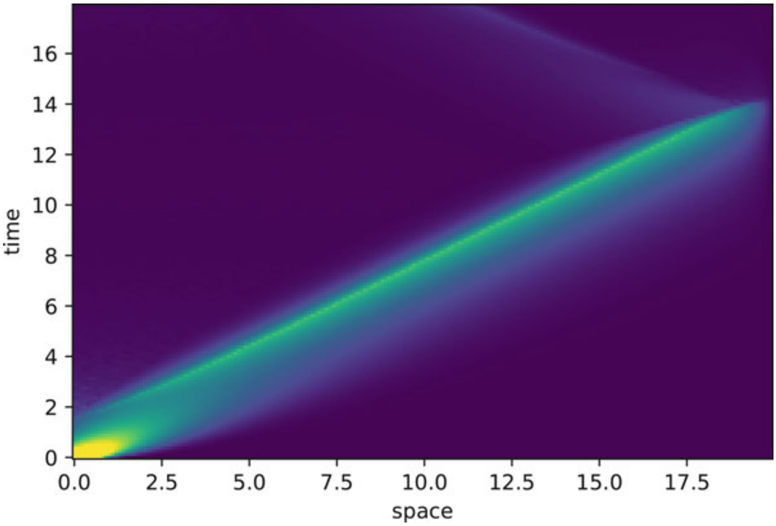


Fig. 5 Numerical simulation of model (3.1) for a half-Gaussian initial density of bacteria

in developmental biology [22], let us point to the modeling study [63], which is devoted to the migration of cell collectives in the lateral line during development of the zebrafish. There, it is assumed that the rod of cells maintains its shape per se with a constant length, which is a parameter of the model; see also [12] for biological evidence of cell attraction during collective motion.

4 Scenario 2: Cell Leakage Compensated by Growth

In this section, we present a recent model of SGG, including localized (signal-dependent) growth [16]. This work was motivated by aerotactic waves of Dicty observed in vertically confined assays, in which oxygen is consumed by the cells and is soon limited at the center of the colony; see Fig. 6. We refer to [16] for the experimental details. The model introduced in [16] was referred to as a "go-or-grow" model, a term coined in a previous work by Hatzikirou et al. [30] in the context of modeling cell invasion in brain tumors. There, the basic hypothesis was that cells could switch between two states, or phenotypes: a migrating state "go" (with enhanced random diffusion), and a proliferating state "grow" (with enhanced rate

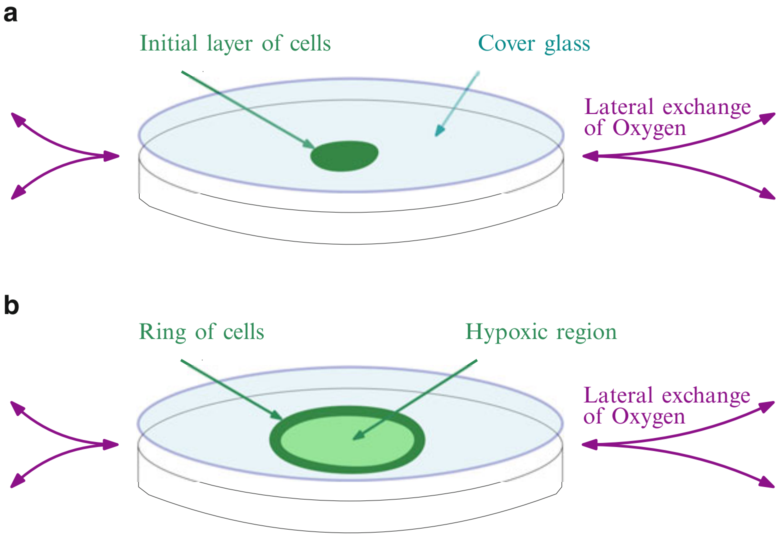


Fig. 6 Schematic view of the experimental setup in [16]. (**a**) An initial layer of Dicty cells is deposited at the center of the plate and covered with a large glass coverslip (after [21]). This vertical confining reduces drastically the inflow of oxygen within the plate, by restricting it to lateral exchanges. (**b**) Soon after the beginning of the experiment, a ring of cells emerges, which is traveling over several days at constant speed with a well-preserved shape. The moving ring consumes almost all the available oxygen, so that the center of the colony is at very low concentration, below 1%

of division), following previous works in the same context (see, e.g., [25]). In [30] it was assumed that hypoxia (lack of oxygen) triggers the switch in the long-term dynamics of the system, by selection of the migrating phenotype, but in a global manner (oxygen supply was accounted for via the constant carrying capacity, as one parameter of the cellular automaton). Later contributions considered PDE models with density-dependent switch (see [62], as opposed to [25] where the switching rate is not modulated, and also the experimental design of density-dependent motility in bacteria [44]).

In [16], the go-or-grow hypothesis was revisited, by studying an expanding ring of Dicty cells, with limited supply of oxygen. Figure 7a shows the cell density profile, as it is observed in experiments. Figure 7b summarizes the minimal assumption of an oxygen-dependent switch, as proposed in [16]. It was hypothesized that the transition between the proliferating state and the migrating state is modulated by the level of oxygen, with a sudden change of phenotype at some threshold S_0. Above this threshold, when oxygen is available in sufficient quantity, cells exhibit slow random (diffusive) motion and divide at some constant rate. Below this threshold, when oxygen is limited, cells stop dividing and move preferentially up the oxygen gradient. The latter hypothesis (directional motion) is different from the aforementioned go-or-grow models [25, 30, 62]. It is consistent with the observations of individual tracking within the cell population in the bulk of the wave in [16].

The following model recapitulates these assumptions:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}\left(-d\dfrac{\partial \rho}{\partial x} + \rho \boldsymbol{\chi}\left(S, \dfrac{\partial S}{\partial x}\right)\right) = \mathbf{r}(S)\rho\,, \\ \dfrac{\partial S}{\partial t} = D\dfrac{\partial^2 S}{\partial x^2} - \mathbf{k}(S,\rho)\,. \end{cases} \tag{4.1}$$

with the specific choice

$$\boldsymbol{\chi}\left(S, \frac{\partial S}{\partial x}\right) = \chi \operatorname{sign}\left(\frac{\partial S}{\partial x}\right) \mathbf{1}_{S<S_0}\,, \quad \mathbf{r}(S) = r\mathbf{1}_{S>S_0}\,. \tag{4.2}$$

This can be viewed as another variation of (2.5) including growth. It can also be viewed as an extension of the celebrated F/KPP equation, with a signal-dependent growth saturation, and including advection (we refer to [54, 11, 76] and references therein for more classical synthesis of the F/KPP equation and the Keller–Segel model of cellular aggregation). Interestingly, an analogous model was proposed in [28], following a general motivation, and beginning with the statement that proliferation is necessary to sustain wave propagation. As compared with (4.1)–(4.2), in the latter work, the reproduction rate $\mathbf{r}$ is signal dependent with a linear dependency, and there is no threshold on the chemosensitivity $\boldsymbol{\chi}$, which is simply a linear function of the gradient $\frac{\partial S}{\partial x}$. As a consequence, the wave speed cannot be calculated analytically, in contrast with (4.1)–(4.2) (see Theorem 4.1 below).

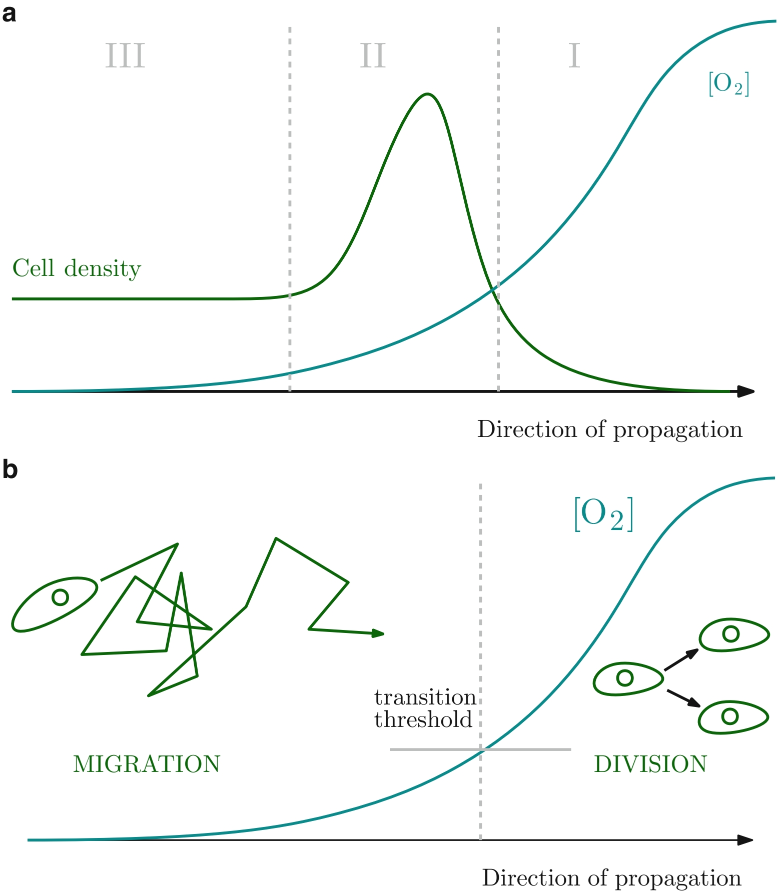


Fig. 7 Graphical description of the "go-or-grow" model (4.1). (**a**) Individual cell tracking in [16] shows different cell behaviors depending on the relative position with respect to the tip of the ring: (I) ahead of the moving ring, cells exhibit unbiased motion, together with division events; (II) inside the ring, cells exhibit clear directional motion (which indeed results in the formation and maintenance of the ring); (III) the trail of cells that are left behind exhibit unbiased motion, again, with more persistent trajectories (but this last observation is neglected in the model, because it was shown to have limited effect). (**b**) We hypothesize a single transition threshold S_0 such that cells can divide above the threshold while they move preferentially up the gradient below the threshold, when oxygen is limited. The unbiased component of cell motion (diffusion) is common to both sides of the threshold

Before we show the construction of traveling wave solutions for (4.1)–(4.2), let us comment on the reason why such solutions can exist. The expected density profile exhibits a plateau of cells left behind the wave; see Fig. 7a. In the vertical confining assay experiment with Dicty, this corresponds to cells that are still highly motile but have lost the propension to move directionally. They cannot keep pace with the self-generated oxygen gradient. The increasing amount of cells that are left behind is compensated by the growth at the edge of the pulse. This localized growth term (above the oxygen threshold) creates a flux term (negative flux in the moving coordinate), which is key to the mathematical construction of the wave.

We can be more precise about the negative flux issued from cell division by looking at the traveling wave equation (4.1)–(4.2) in the moving coordinate $z = x - ct$.

$$-c\frac{d\rho}{dz} + \frac{d}{dz}\left(-d\frac{d\rho}{dz} + \rho\boldsymbol{\chi}\left(S, \frac{dS}{dz}\right)\right) = \mathbf{r}(S)\rho\,. \tag{4.3}$$

Below the oxygen threshold, $S < S_0$, the right-hand side vanishes, and we are left with a constant flux:

$$-c\rho - d\frac{d\rho}{dz} + \rho\boldsymbol{\chi}\left(S, \frac{dS}{dz}\right) = -J\,. \tag{4.4}$$

By integrating (4.3) on $\{S > S_0\}$, and using the continuity of the flux at the interface $\{S = S_0\}$, we find

$$J = r\int_{\{S>S_0\}} \rho(z)\,dz\,. \tag{4.5}$$

Note that the continuity of the flux is a prerequisite for the well-posedness of (4.1)–(4.2); see [20] for a rigorous analysis of this problem, and unexpected mathematical subtleties.

Theorem 4.1 (Cochet et al. [16], Demircigil [20]) *There exist a speed $c > 0$ and a positive limit value $\rho_- > 0$, such that the system (4.1)–(4.2) admits a stationary solution in the moving frame $(\rho(x - ct), S(x - ct))$, such that ρ and S have the following limiting values:*

$$\begin{cases}\lim_{z\to-\infty}\rho(z) = \rho_-\,,\\ \lim_{z\to+\infty}\rho(z) = 0\,,\end{cases} \qquad \begin{cases}\lim_{z\to-\infty}S(z) = 0\,,\\ \lim_{z\to+\infty}S(z) = S_{\text{init}}\,.\end{cases}$$

Moreover, the speed is given by the following dichotomy:

$$c = \begin{cases}2\sqrt{rd} & \text{if } \chi \le \sqrt{rd}\,,\\ \chi + \dfrac{rd}{\chi} & \text{if } \chi \ge \sqrt{rd}\,.\end{cases} \tag{4.6}$$

Interestingly, the dichotomy in (4.6) depends on the relative values of the advection speed (up the gradient) χ, and half the reaction–diffusion speed of the F/KPP equation $\sqrt{rd}$. When the aerotactic biases are small (low advection speed χ), then the wave is essentially driven by growth and diffusion. When biases are large, then the wave is mainly driven by aerotaxis. This has interesting implications in terms of maintenance of genetic diversity inside the wave (see [6, 52] for diversity dynamics among reaction–diffusion traveling waves). In fact, the so-called dichotomy between *pulled* and *pushed* waves is at play here; see [16, 20] for more details and discussion.

In contrast to the original Keller–Segel model (2.2), the wave speed does not depend on the features of oxygen consumption and diffusion.

Proof As in Sect. 3, the wave speed is not given a priori. We seek a monotonic oxygen profile, such that $\frac{dS}{dz} > 0$. Therefore, the first equation reduces to

$$-c\frac{d\rho}{dz} - d\frac{d^2\rho}{dz^2} + \frac{d}{dz}\left\{\begin{matrix} 0 & \text{if } S > S_0 \\ \chi\rho & \text{if } S < S_0 \end{matrix}\right\} = \left\{\begin{matrix} r\rho & \text{if } S > S_0 \\ 0 & \text{if } S < S_0 \end{matrix}\right\}.$$

By translational invariance, we assume that $S = S_0$ occurs at $z = 0$.

For $z < 0$, we have by (4.4)–(4.5),

$$d\frac{d\rho}{dz} = J + (\chi - c)\rho\,, \quad J > 0\,. \tag{4.7}$$

Suppose that $c \leq \chi$. Then, $d\frac{d\rho}{dz} \geq J > 0$, which is a contradiction with the positivity of ρ. Hence, we must have $c > \chi$. The solution of (4.7) is unbounded unless it is constant, that is $\rho = \frac{J}{c-\chi}$, and this is the natural choice we make for the construction.

For $z > 0$, we have the standard linear problem arising in the F/KPP equation (at small density):

$$-c\frac{d\rho}{dz} - d\frac{d^2\rho}{dz^2} = r\rho\,.$$

We look for exponential solutions $\exp(-\lambda z)$. The characteristic equation $d\lambda^2 - c\lambda + r = 0$ has real roots when $c^2 \geq 4rd$. Then, we proceed by dichotomy.

- ◇ The case $c = 2\sqrt{rd}$. The general solution for $z > 0$ is of the form $(a + bz)\exp(-\lambda z)$, with $\lambda = \sqrt{\frac{r}{d}}$ the double root. The constant a coincides with $\frac{J}{c-\chi}$ by continuity of the density (its value does not really matter here). Continuity of the flux at the interface $z = 0$ yields $-d(b - a\lambda) = \chi a$, hence $bd = a(\sqrt{rd} - \chi)$. Thus, the solution is admissible ($b \geq 0$) if and only if $\chi \leq \sqrt{rd}$.
- ◇ The case $c > 2\sqrt{rd}$. Standard arguments in the construction of reaction–diffusion traveling waves imply to select the sharpest decay on the right side [3, 71], namely $\rho = a\exp(-\lambda z)$, with $\lambda = \frac{1}{2d}\left(c + \sqrt{c^2 - 4rd}\right)$. Continuity of the flux at the interface now writes $-d(-a\lambda) = \chi a$, which is equivalent to

$$2\chi - c = \sqrt{c^2 - 4rd} \quad \Leftrightarrow \quad \left(c = \chi + \frac{rd}{\chi}\right) \;\&\; \left(\chi > \frac{c}{2}\right).$$

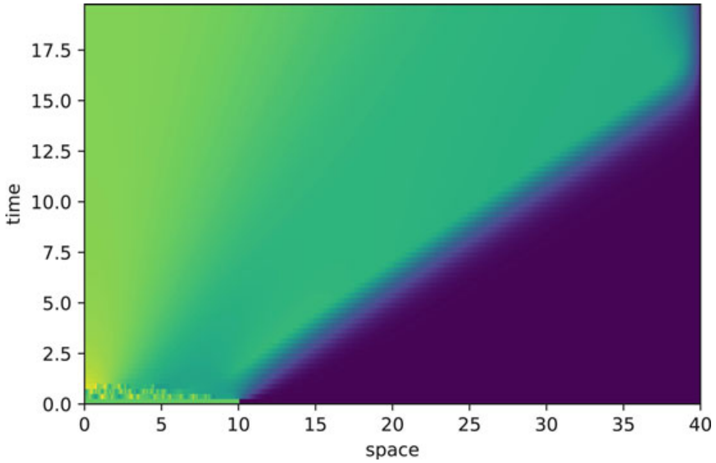


Fig. 8 Numerical simulation of model (3.1) for an initial plateau of cells restricted to the interval $\{x < 10\}$

It must be checked a posteriori that $c > 2\sqrt{rd}$, which is immediate. The last inequality constraint ensures that $\chi > \sqrt{rd}$, in contrast to the other side of the dichotomy.

Thus, the construction is complete. □

The wavefront constructed above appears to be numerically stable, driving the long-time asymptotics; see Fig. 8. However, the very strong advection at the back of the wave creates a decreasing density profile, which is actually constant at the back of the wavefront, in contrast to the experiments showing a non-monotonic pulse (Fig. 7). Several extensions were discussed in [16].

Logarithmic Sensitivity Below, we discuss a natural, yet original, extension of the previous result, restoring the logarithmic gradient in the advection term. More precisely, we consider (4.1) again, with the following choice of functions, instead of (4.2):

$$\boldsymbol{\chi}\left(S, \frac{\partial S}{\partial x}\right) = \chi \log\left(\frac{\partial S}{\partial x}\right) \mathbf{1}_{S<S_0}\,, \quad \mathbf{r}(S) = r\mathbf{1}_{S>S_0}\,. \tag{4.8}$$

We present below a preliminary result about the existence of traveling waves, followed by heuristic arguments about the determination of the speed, and some numerical investigation.

Theorem 4.2 *Assume $D = 0$, and $\mathbf{k}(S, \rho) = k\rho S$ for some $k > 0$. There exist a speed $c > 0$ and a positive limit value $\rho_- > 0$, such that the system (4.1)–(4.8) admits a stationary solution in the moving frame $(\rho(x - ct), S(x - ct))$, such that ρ and S have the following limiting values:*

$$\begin{cases}\lim_{z\to-\infty}\rho(z)=\rho_-\,, \\ \lim_{z\to+\infty}\rho(z)=0\,,\end{cases}\qquad \begin{cases}\lim_{z\to-\infty}S(z)=0\,, \\ \lim_{z\to+\infty}S(z)=S_{\mathrm{init}}\,.\end{cases}$$

Moreover, the speed is given by the following dichotomy:

$$c=2\sqrt{r\max\left\{d,\,\chi\log\left(\frac{S_{\mathrm{init}}}{S_0}\right)\right\}}\,. \tag{4.9}$$

Proof We proceed similarly as in the proof of the previous statement. The assumption $D=0$ enables expressing the logarithmic gradient in terms of the density:

$$-c\frac{d(\log S)}{dz}=-k\rho\,. \tag{4.10}$$

For $z<0$, we have a constant (negative) flux at equilibrium in the moving frame (4.4):

$$-c\rho-d\frac{d\rho}{dz}+\chi\rho\frac{d(\log S)}{dz}=-J<0\,. \tag{4.11}$$

Combining (4.10) and (4.11), we get the ODE satisfied by the cell density profile at the back:

$$d\frac{d\rho}{dz}=-c\rho+\frac{k\chi}{c}\rho^2+J\,. \tag{4.12}$$

This ODE comes with a sign condition, for the discriminant of the right-hand side to be nonnegative (otherwise ρ cannot be positive for all $z<0$ when $\frac{d\rho}{dz}$ is uniformly positive), that is,

$$\frac{c^3}{4k\chi}\geq J\,. \tag{4.13}$$

This condition is complemented by the integration of (4.10) over $\{z>0\}$:

$$c\log\left(\frac{S_{\mathrm{init}}}{S_0}\right)=k\int_0^{+\infty}\rho(z)\,dz=\frac{k}{r}J\,,$$

where the last identity follows from (4.5). This yields the constraint

$$\frac{c^3}{4r\chi}\geq c\log\left(\frac{S_{\mathrm{init}}}{S_0}\right)\quad\Leftrightarrow\quad c^2\geq 4r\chi\log\left(\frac{S_{\mathrm{init}}}{S_0}\right)\,. \tag{4.14}$$

This is one part of the condition in (4.9). The second part comes naturally from the constraint on the characteristic equation on $\{z>0\}$, namely $c^2\geq 4rd$. It can be shown by simple phase plane analysis that admissible solutions exist in both cases when the inequality (4.9) is an equality. □

The previous analysis calls for a few comments:

1. Contrary to the former construction in Theorem 4.1, the latter construction does not come naturally with an equation for c. This is because there is no clear way to remove one degree of freedom on $\{z < 0\}$ under the sign condition (4.13). Indeed, the solution of (4.12) is naturally bounded for any initial condition, in opposition to (4.7).
2. Surprisingly, the additional restriction (4.14) results from conditions imposed on the solution *at the back of the wave* on $\{z < 0\}$, in opposition to the standard case, say for F/KPP and related equations, where it always comes from conditions on $\{z > 0\}$ (as it is the case for the classical restriction $c^2 \geq 4rd$).

At this point, we conjecture that the minimal speed (4.9) giving rise to admissible solutions is selected when the Cauchy problem is initiated with localized initial data.

Claim 4.3 Starting from a compactly supported initial data, the asymptotic spreading speed of solutions to (4.1)–(4.8) is given by (4.9).

This claim is supported by numerical exploration of the system in some range of parameters; see Fig. 9 for one typical set of parameters. On the one hand, the claim is not surprising in the case of small bias, when $c = 2\sqrt{rd}$. In fact, this corresponds to the standard mechanism of speed determination at the edge of the front in reaction–diffusion equation with pulled waves. This was indeed confirmed in the previous model (4.1)–(4.2) [16, 20]. On the other hand, we emphasize that it does look surprising in the case of large bias, when $c = 2\sqrt{r\chi \log\left(\frac{S_{\text{init}}}{S_0}\right)}$. In the latter case, the selection of the minimal speed would come from a discriminant condition *at the back of the wave*, which would be a quite original phenomenon, up to our knowledge.

5 Conclusion and Perspectives

We exposed the original contribution of Keller and Segel devoted to chemotactic waves of bacteria and discussed its limitations. These limitations are mainly concerned with the possible lack of positivity of the chemical concentration in the model. A pair of extensions were described. They both resolve the positivity issue while keeping analytical solvability of the waves, thanks to the specific choice of piecewise homogeneous models. In addition, they are both supported by biological experiments, respectively, with bacteria *E. coli* and Dicty cells.

To conclude, let us mention some open problems, either on the mathematical or on the modeling side.

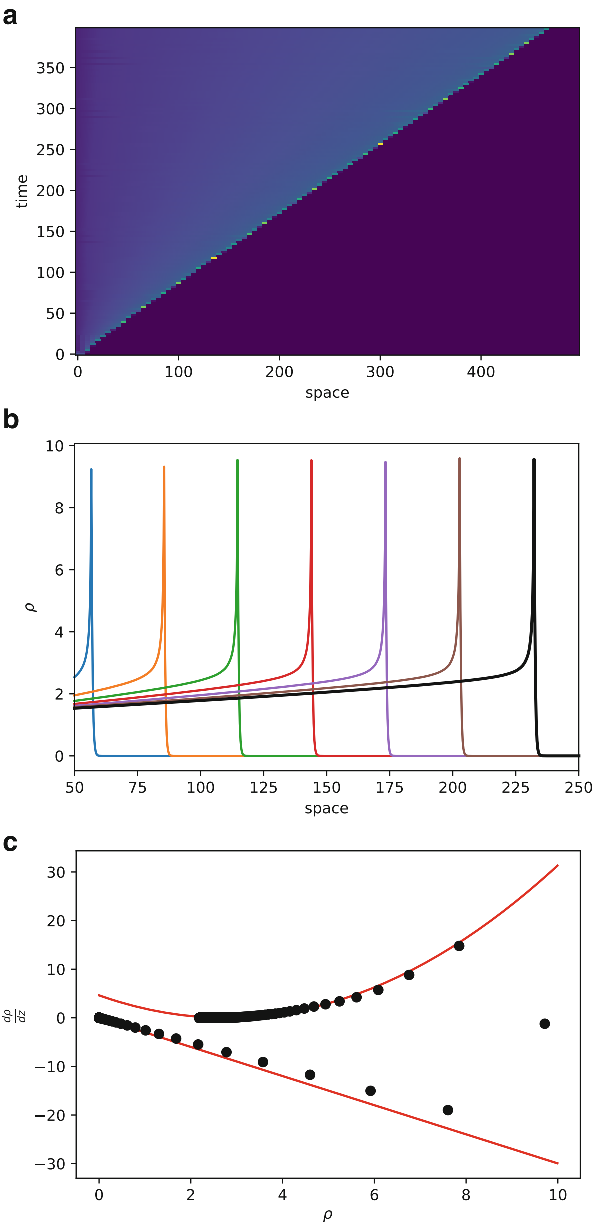


Fig. 9 (**a**) Traveling wave propagation obtained after long time simulations of the Cauchy problem (4.1)–(4.8) with parameters ($d = 1$, $\chi = 2$, $r = 1$, $D = 0$, $k = 1$, $S_{\text{init}} = 8$, $S_0 = 2$).

Determinacy of the Speed at the Back of the Wave The result stated in Theorem 4.2 appeared quite unexpectedly. If further numerical exploration with alternative schemes tends to confirm our Claim 4.3, we believe that understanding the mechanism of speed selection is an interesting and possibly original problem per se. We stress out that this mechanism occurs at $z = -\infty$, in the sense that the sign condition on the discriminant in (4.12) ensures that the cell density remains positive for negative z. Alternatively speaking, we face a situation that is the mirror of the standard mechanism of speed determinacy at $z = +\infty$ in the F/KPP equation.

Traveling Waves with Nonzero Chemical Diffusion Figure 10 shows the numerical simulation of the Cauchy problem (4.1)–(4.8) with a chemical diffusion coefficient D of order one. It seems that the solution converges toward a traveling wave profile as $t \to +\infty$ with reduced speed as compared to the case without chemical diffusion (Fig. 9). Moreover, the numerical wave plotted in the phase plane shows a similar pattern (compare Fig. 9c, b), suggesting similar mechanisms occurring at $z = -\infty$ (in particular, a vanishing discriminant in the super-critical case $c > 2\sqrt{rd}$). However, since the relationship (4.10) is not satisfied with nonzero diffusion, we are lacking one equation to perform explicit computations. There exist multiple works extending the construction of waves for the original model (2.1) to the case of nonzero chemical diffusion. This may give some hints to address this question.

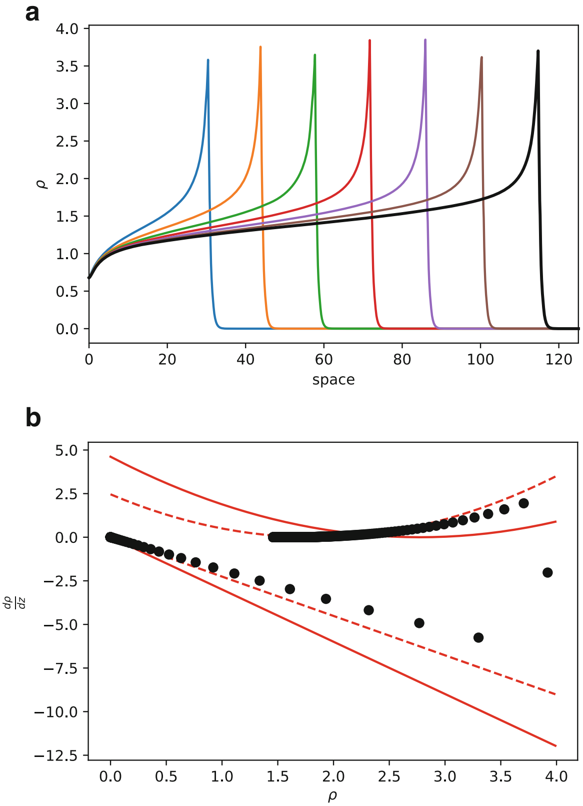
Stability Although stability in the Keller–Segel model (2.1) has drawn some attention, with a nearly complete picture by now, stability of the traveling wave solutions to the models presented in Sects. 3 and 4 is almost entirely open. The first author and Hoffmann proved local non-linear stability of standing waves for (3.1) (without the SGG signaling S), assuming that the attractant concentration A is quasi-stationary (solving an elliptic equation at any time). They performed a change of coordinates to bypass the discontinuity of the advection coefficient and used higher-order energy methods to handle the singular term of the coupling.

Nevertheless, numerical investigation performed at the occasion of this work, with simple finite volume, semi-implicit, upwind schemes, argues in favor of stability of all the waves described in Sects. 3 and 4.

Spatial Sorting Another open problem is the theoretical analysis of spatial sorting in bacteria collectives when the individuals have different chemotactic sensitivities.

Fig. 9 (continued) (**b**) The density profile is shown at successive times in the moving frame. Note the low decay at the back of the wave, which is the signature of singular point in the ODE (4.12) together with the choice of J that cancels the discriminant in (4.13). The numerical speed is $c_{\text{num}} \approx 3.17$, close to the theoretical one, $2\sqrt{\log(4)} \approx 3.33$. (**c**) To better assess our Claim 4.3, the numerical solution is plotted in the phase plane (ρ, ρ') (black dots), against the theoretical curves, that is, $\rho' = -\lambda\rho$ (for $z > 0$), and $\rho' = \frac{k\chi}{cd}\left(\rho - \frac{c^2}{2k\chi}\right)^2$ (4.12) (red lines). The isolated point on the right corresponds to the transition at $z = 0$, where the expected theoretical profile has a $\mathcal{C}^1$ discontinuity. We believe that the discrepancy is due to numerical errors

Fig. 10 Same as in Fig. 9, except for the diffusion coefficient of the chemical that is set to $D = 1$. (**a**) We observe propagation of a traveling wave in the long time asymptotic with a reduced speed. Clearly, the wave profile differs significantly from Fig. 9b. (**b**) In particular, the solution in the phase plane does not align with the theoretical expectation available in the case $D = 0$ (red plain curves). It aligns much better with the theoretical expectation computed from Eqs. (4.10)–(4.14) taking the reduced numerical speed as an input (red dashed curves). We believe that the discrepancy is due to numerical errors



In [29], remarkable experiments on bacteria *E. coli*, together with a very elegant analytical argument, indicated that cells can move together despite their differences. The argument of [29] goes as follows: assume that there exist multiple types of bacteria consuming a single nutrient S and that each type is characterized by a chemotactic sensitivity χ_i; suppose that, for each type, the chemotactic advection is of the form $\boldsymbol{\chi}_i\left(S, \frac{\partial S}{\partial x}\right) = \chi_i \frac{\partial F(S)}{\partial x}$, say the logarithmic gradient as in the original model (2.1); suppose that the solution of each type converges toward a traveling wave in the long time, with a common speed c, so that the flux is asymptotically zero in the moving frame for each type:

$$(\forall i) \qquad -c - d\frac{d}{dz}(\log \rho_i) + \chi_i \frac{d}{dz}F(S) = 0\,. \tag{5.1}$$

Evaluating (5.1) at the maximum point of the density ρ_i, say z_i^*, we would get that

$$c = \chi_i \frac{d}{dz}F(S)(z_i^*)\,. \tag{5.2}$$

Differentiating (5.1) at $z = z_i^*$, it could be deduced that

$$\frac{d^2}{dz^2} F(S)(z_i^*) = d \frac{d^2}{dz^2} (\log \rho_i) (z_i^*) \leq 0 . \tag{5.3}$$

The combination of (5.2) and (5.3) says that the peaks (z_i^*) of the densities (ρ_i) that are traveling together are restricted to the interval where $F(S)$ is concave. Moreover, they are ordered in such a way that $(\chi_i < \chi_j) \Rightarrow (z_i^* < z_j^*)$. This nice calculation indicates that different phenotypes could migrate collectively despite their differences. The intuitive reason, which can be read on (5.2), is that larger chemosensitivity χ_i naturally pushes the cells ahead, where they experience shallower gradients. Nonetheless, the analysis in [29] is not complete, as the existence of stable traveling waves of different types with a common speed is taken for granted.

There exist previous theoretical works about collective migration of different phenotypes within the same chemical environment. We refer, for instance, to [43], which adopted the framework of the original model by Keller and Segel (2.1). In view of the discussion above, the stability of their theoretical outcomes is questionable. In [23], the authors extend the framework of Sect. 3, including two subpopulations with different chemotactic phenotypes. This work was supported by experimental data. However, the discussion in [29] makes it clear that the framework of [23] is not directly compatible with their findings. Actually, it is one consequence of the advection speed discontinuity in (3.1) that the maximum peak density is located at the sign transition, whatever the chemosensitivity coefficient is, hence violating the nice relationship (5.2).

Preliminary investigations suggest that the framework of Sect. 4 cannot be readily extended as well. Indeed, signal-dependent growth counterbalances the fact that more efficient chemotactic types experience shallower gradients, because they have better access to nutrient. This triggers natural selection of the more efficient type by differential growth (results not shown).

To our knowledge, there is no clear mathematical framework to handle the remarkable experiments and biological insights as shown in [29], at the present time.

Acknowledgments Part of this work has been achieved during the third author's master internship at Institut Camille Jordan. The two first authors are indebted to Christophe Anjard, Olivier Cochet-Escartin, and Jean-Paul Rieu for having drawn their attention to SGG beyond the case of bacterial waves. The authors are very grateful to Eitan Tadmor for his constant encouragement to write this contribution, and to a pair of anonymous reviewers for their feedbacks. This project has received financial support from the CNRS through the MITI interdisciplinary programs. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 865711).

References

1. J. Adler. Chemotaxis in Bacteria. *Science*, 153(3737):708–716, Dec. 1966.
2. W. Alt. Biased random walk models for chemotaxis and related diffusion approximations. *Journal of Mathematical Biology*, 9(2):147–177, Apr. 1980.
3. D. G. Aronson and H. F. Weinberger. Multidimensional nonlinear diffusion arising in population genetics. *Advances in Mathematics*, 30(1):33–76, Oct. 1978.
4. N. Bellomo, A. Bellouquid, Y. Tao, and M. Winkler. Toward a mathematical theory of Keller–Segel models of pattern formation in biological tissues. *Mathematical Models and Methods in Applied Sciences*, 25(09):1663–1763, Mar. 2015.
5. H. C. Berg. *E. coli in motion*. Springer, 2004.
6. G. Birzu, O. Hallatschek, and K. S. Korolev. Fluctuations uncover a distinct class of traveling waves. *Proceedings of the National Academy of Sciences*, 115(16):E3645–E3654, Apr. 2018.
7. M. P. Brenner, L. S. Levitov, and E. O. Budrene. Physical Mechanisms for Chemotactic Pattern Formation by Bacteria. *Biophysical Journal*, 74(4):1677–1693, Apr. 1998.
8. E. O. Budrene and H. C. Berg. Complex patterns formed by motile cells of Escherichia coli. *Nature*, 349(6310):630–633, Feb. 1991.
9. E. O. Budrene and H. C. Berg. Dynamics of formation of symmetrical patterns by chemotactic bacteria. *Nature*, 376(6535):49–53, July 1995.
10. V. Calvez. Chemotactic waves of bacteria at the mesoscale. *Journal of the European Mathematical Society*, 22(2):593–668, Nov. 2019.
11. V. Calvez, B. Perthame, and S. Yasuda. Traveling wave and aggregation in a flux-limited Keller-Segel model. *Kinetic & Related Models*, 11(4):891, 2018.
12. C. Carmona-Fontaine, E. Theveneau, A. Tzekou, M. Tada, M. Woods, K. M. Page, M. Parsons, J. D. Lambris, and R. Mayor. Complement Fragment C3a Controls Mutual Cell Attraction during Collective Cell Migration. *Developmental Cell*, 21(6):1026–1037, Dec. 2011.
13. J. A. Carrillo, J. Li, and Z.-A. Wang. Boundary spike-layer solutions of the singular Keller–Segel system: existence and stability. *Proceedings of the London Mathematical Society*, 122(1):42–68, 2021.
14. F. Chalub, Y. Dolak-Struss, P. Markowich, D. Oelz, C. Schmeiser, and A. Soreff. Model hierarchies for cell aggregation by chemotaxis. *Mathematical Models and Methods in Applied Sciences*, 16(supp01):1173–1197, July 2006.
15. F. A. C. C. Chalub, P. A. Markowich, B. Perthame, and C. Schmeiser. Kinetic Models for Chemotaxis and their Drift-Diffusion Limits. *Monatshefte für Mathematik*, 142(1–2):123–141, June 2004.
16. O. Cochet-Escartin, M. Demircigil, S. Hirose, B. Allais, P. Gonzalo, I. Mikaelian, K. Funamoto, C. Anjard, V. Calvez, and J.-P. Rieu. Hypoxia triggers collective aerotactic migration in Dictyostelium discoideum. *eLife*, 10:e64731, Aug. 2021.
17. J. Cremer, T. Honda, Y. Tang, J. Wong-Ng, M. Vergassola, and T. Hwa. Chemotaxis as a navigation strategy to boost range expansion. *Nature*, 575(7784):658–663, Nov. 2019.
18. P. N. Davis, P. v. Heijster, and R. Marangell. Absolute instabilities of travelling wave solutions in a Keller–Segel model. *Nonlinearity*, 30(11):4029–4061, Oct. 2017.
19. P. N. Davis, P. van Heijster, and R. Marangell. Spectral stability of travelling wave solutions in a Keller–Segel model. *Applied Numerical Mathematics*, 141:54–61, July 2019.
20. M. Demircigil. *in preparation*.
21. M. Deygas, R. Gadet, G. Gillet, R. Rimokh, P. Gonzalo, and I. Mikaelian. Redox regulation of EGFR steers migration of hypoxic mammary cells towards oxygen. *Nature Communications*, 9(1):4545, Oct. 2018.
22. E. Donà, J. D. Barry, G. Valentin, C. Quirin, A. Khmelinskii, A. Kunze, S. Durdu, L. R. Newton, A. Fernandez-Minan, W. Huber, M. Knop, and D. Gilmour. Directional tissue migration through a self-generated chemokine gradient. *Nature*, 503(7475):285–289, Nov. 2013.

23. C. Emako, C. Gayrard, A. Buguin, L. N. d. Almeida, and N. Vauchelet. Traveling Pulses for a Two-Species Chemotaxis Model. *PLOS Comput Biol*, 12(4):e1004843, Apr. 2016.
24. R. Erban and H. G. Othmer. From Individual to Collective Behavior in Bacterial Chemotaxis. *SIAM Journal on Applied Mathematics*, 65(2):361–391, Jan. 2004.
25. S. Fedotov and A. Iomin. Migration and proliferation dichotomy in tumor cell invasion. *Physical Review Letters*, 98(11):118101, Mar. 2007. arXiv: q-bio/0610016.
26. F. Filbet, P. Laurençot, and B. Perthame. Derivation of hyperbolic models for chemosensitive movement. *Journal of Mathematical Biology*, 50(2):189–207, Feb. 2005.
27. R. A. Fisher. The Wave of Advance of Advantageous Genes. *Annals of Eugenics*, 7(4):355–369, 1937.
28. B. Franz, C. Xue, K. J. Painter, and R. Erban. Travelling Waves in Hybrid Chemotaxis Models. *Bulletin of Mathematical Biology*, 76(2):377–400, Dec. 2013.
29. X. Fu, S. Kato, J. Long, H. H. Mattingly, C. He, D. C. Vural, S. W. Zucker, and T. Emonet. Spatial self-organization resolves conflicts between individuality and collective migration. *Nature Communications*, 9(1):2177, June 2018.
30. H. Hatzikirou, D. Basanta, M. Simon, K. Schaller, and A. Deutsch. 'Go or Grow': the key to the emergence of invasion in tumour progression? *Mathematical Medicine and Biology*, 29(1):49–65, Mar. 2012.
31. T. Hillen and K. J. Painter. A user's guide to PDE models for chemotaxis. *Journal of Mathematical Biology*, 58(1–2):183–217, Jan. 2009.
32. M. Hilpert. Lattice-Boltzmann model for bacterial chemotaxis. *Journal of Mathematical Biology*, 51(3):302–332, Sept. 2005.
33. D. Horstmann and A. Stevens. A Constructive Approach to Traveling Waves in Chemotaxis. *Journal of Nonlinear Science*, 14(1):1–25, Jan. 2004.
34. L. Jiang, Q. Ouyang, and Y. Tu. Quantitative Modeling of Escherichia coli Chemotactic Motion in Environments Varying in Space and Time. *PLOS Comput Biol*, 6(4):e1000735, Apr. 2010.
35. H.-Y. Jin, J. Li, and Z.-A. Wang. Asymptotic stability of traveling waves of a chemotaxis model with singular sensitivity. *Journal of Differential Equations*, 255(2):193–219, July 2013.
36. Y. V. Kalinin, L. Jiang, Y. Tu, and M. Wu. Logarithmic Sensing in Escherichia coli Bacterial Chemotaxis. *Biophysical Journal*, 96(6):2439–2448, Mar. 2009.
37. E. F. Keller and L. A. Segel. Conflict between Positive and Negative Feedback as an Explanation for the Initiation of Aggregation in Slime Mould Amoebae. *Nature*, 227(5265):1365–1366, Sept. 1970.
38. E. F. Keller and L. A. Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26(3):399–415, Mar. 1970.
39. E. F. Keller and L. A. Segel. Model for chemotaxis. *Journal of Theoretical Biology*, 30(2):225–234, Feb. 1971.
40. E. F. Keller and L. A. Segel. Traveling bands of chemotactic bacteria: A theoretical analysis. *Journal of Theoretical Biology*, 30(2):235–248, Feb. 1971.
41. A. N. Kolmogorov, I. G. Petrovsky, and N. S. Piskunov. Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Mosc. Univ. Bull. Math*, 1:1–25, 1937.
42. J. Li, T. Li, and Z.-A. Wang. Stability of traveling waves of the Keller–Segel system with logarithmic sensitivity. *Mathematical Models and Methods in Applied Sciences*, 24(14):2819–2849, Dec. 2014.
43. T.-C. Lin, and Z.-A. Wang. Development of traveling waves in an interacting two-species chemotaxis model. *Discrete & Continuous Dynamical Systems - A*, 34(7):2907–2927, 2014.
44. C. Liu, X. Fu, L. Liu, X. Ren, C. K. L. Chau, S. Li, L. Xiang, H. Zeng, G. Chen, L.-H. Tang, P. Lenz, X. Cui, W. Huang, T. Hwa, and J.-D. Huang. Sequential Establishment of Stripe Patterns in an Expanding Cell Population. *Science*, 334(6053):238–241, Oct. 2011.
45. R. Majumdar, M. Sixt, and C. A. Parent. New paradigms in the establishment and maintenance of gradients during directed cell migration. *Current Opinion in Cell Biology*, 30:33–40, Oct. 2014.

46. N. Mittal, E. O. Budrene, M. P. Brenner, and A. v. Oudenaarden. Motility of Escherichia coli cells in clusters formed by chemotactic aggregation. *Proceedings of the National Academy of Sciences*, 100(23):13259–13263, Nov. 2003.
47. A. J. Muinonen-Martin, O. Susanto, Q. Zhang, E. Smethurst, W. J. Faller, D. M. Veltman, G. Kalna, C. Lindsay, D. C. Bennett, O. J. Sansom, R. Herd, R. Jones, L. M. Machesky, M. J. O. Wakelam, D. A. Knecht, and R. H. Insall. Melanoma Cells Break Down LPA to Establish Local Gradients That Drive Chemotactic Dispersal. *PLOS Biology*, 12(10):e1001966, Oct. 2014.
48. T. Nagai and T. Ikeda. Traveling waves in a chemotactic model. *Journal of Mathematical Biology*, 30(2):169–184, Nov. 1991.
49. H. G. Othmer, S. R. Dunbar, and W. Alt. Models of dispersal in biological systems. *Journal of Mathematical Biology*, 26(3):263–298, June 1988.
50. H. G. Othmer and T. Hillen. The Diffusion Limit of Transport Equations II: Chemotaxis Equations. *SIAM Journal on Applied Mathematics*, 62(4):1222–1250, Jan. 2002.
51. M. A. Rivero, R. T. Tranquillo, H. M. Buettner, and D. A. Lauffenburger. Transport models for chemotactic cell populations based on individual cell behavior. *Chemical Engineering Science*, 44(12):2881–2897, Jan. 1989.
52. L. Roques, J. Garnier, F. Hamel, and E. K. Klein. Allee effect promotes diversity in traveling waves of colonization. *Proceedings of the National Academy of Sciences*, 109(23):8828–8833, May 2012.
53. G. Rosen and S. Baloga. On the stability of steadily propagating bands of chemotactic bacteria. *Mathematical Biosciences*, 24(3):273–279, Jan. 1975.
54. L. Ryzhik, B. Perthame, and G. Nadin. Traveling waves for the Keller–Segel system with Fisher birth terms. *Interfaces and Free Boundaries*, 10(4):517–538, Dec. 2008.
55. B. Sandstede. Chapter 18 - Stability of Travelling Waves. In B. Fiedler, editor, *Handbook of Dynamical Systems*, volume 2 of *Handbook of Dynamical Systems*, pages 983–1055.
56. J. Saragosti, V. Calvez, N. Bournaveas, A. Buguin, P. Silberzan, and B. Perthame. Mathematical Description of Bacterial Traveling Pulses. *PLoS Computational Biology*, 6(8):e1000890, Aug. 2010.
57. J. Saragosti, V. Calvez, N. Bournaveas, B. Perthame, A. Buguin, and P. Silberzan. Directional persistence of chemotactic bacteria in a traveling concentration wave. *Proceedings of the National Academy of Sciences*, 108(39):16235–16240, Sept. 2011.
58. C. Scherber, A. J. Aranyosi, B. Kulemann, S. P. Thayer, O. Iliopoulos, and D. Irimia. Epithelial cell guidance by self-generated EGF gradients. *Integr Biol*, page 22, 2012.
59. N. Sfakianakis, A. Madzvamuse, and M. A. J. Chaplain. A Hybrid Multiscale Model for Cancer Invasion of the Extracellular Matrix. *Multiscale Modeling & Simulation*, 18(2):824–850, Jan. 2020.
60. G. Si, M. Tang, and X. Yang. A Pathway-Based Mean-Field Model for E. coli Chemotaxis: Mathematical Derivation and Its Hyperbolic and Parabolic Limits. *Multiscale Modeling & Simulation*, 12(2):907–926, Jan. 2014.
61. G. Si, T. Wu, Q. Ouyang, and Y. Tu. Pathway-Based Mean-Field Model for *Escherichia coli* Chemotaxis. *Physical Review Letters*, 109(4):048101, July 2012.
62. T. L. Stepien, E. M. Rutter, and Y. Kuang. Traveling Waves of a Go-or-Grow Model of Glioma Growth. *SIAM Journal on Applied Mathematics*, 78(3):1778–1801, Jan. 2018.
63. S. J. Streichan, G. Valentin, D. Gilmour, and L. Hufnagel. Collective cell migration guided by dynamically maintained gradients. *Phys. Biol.*, page 9, 2011.
64. D. W. Stroock. Some stochastic processes which arise from a model of the motion of a bacterium. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 28(4):305–315, Dec. 1974.
65. M. J. Tindall, P. K. Maini, S. L. Porter, and J. P. Armitage. Overview of Mathematical Approaches Used to Model Bacterial Chemotaxis II: Bacterial Populations. *Bulletin of Mathematical Biology*, 70(6):1570–1607, July 2008.
66. Y. Tu, T. S. Shimizu, and H. C. Berg. Modeling the chemotactic response of Escherichia coli to time-varying stimuli. *Proceedings of the National Academy of Sciences*, 105(39):14855–14860, Sept. 2008.

67. L. Tweedy and R. H. Insall. Self-Generated Gradients Yield Exceptionally Robust Steering Cues. *Frontiers in Cell and Developmental Biology*, 8, 2020.
68. L. Tweedy, D. A. Knecht, G. M. Mackay, and R. H. Insall. Self-Generated Chemoattractant Gradients: Attractant Depletion Extends the Range and Robustness of Chemotaxis. *PLOS Biology*, 14(3):e1002404, Mar. 2016.
69. L. Tweedy, O. Susanto, and R. H. Insall. Self-generated chemotactic gradients—cells steering themselves. *Current Opinion in Cell Biology*, 42:46–51, Oct. 2016.
70. L. Tweedy, P. A. Thomason, P. I. Paschke, K. Martin, L. M. Machesky, M. Zagnoni, and R. H. Insall. Seeing around corners: Cells solve mazes and respond at a distance using attractant breakdown. *Science*, 369(6507), Aug. 2020.
71. W. van Saarloos. Front propagation into unstable states. *Physics Reports*, 386(2):29–222, Nov. 2003.
72. G. Venkiteswaran, S. Lewellis, J. Wang, E. Reynolds, C. Nicholson, and H. Knaut. Generation and Dynamics of an Endogenous, Self-Generated Signaling Gradient across a Migrating Tissue. *Cell*, 155(3):674–687, Oct. 2013.
73. Z.-A. Wang. Mathematics of traveling waves in chemotaxis –Review paper–. *Discrete and Continuous Dynamical Systems - Series B*, 18(3):601–641, Dec. 2012.
74. C. Xue. Macroscopic equations for bacterial chemotaxis: integration of detailed biochemistry of cell signaling. *Journal of Mathematical Biology*, 70(1):1–44, Jan. 2015.
75. C. Xue, H. J. Hwang, K. J. Painter, and R. Erban. Travelling Waves in Hyperbolic Chemotaxis Equations. *Bulletin of Mathematical Biology*, 73(8):1695–1733, Aug. 2011.
76. Y. Zeng and K. Zhao. On the logarithmic Keller-Segel-Fisher/KPP system. *Discrete & Continuous Dynamical Systems*, 39(9):5365, 2019.

Clustering Dynamics on Graphs: From Spectral Clustering to Mean Shift Through Fokker–Planck Interpolation

Katy Craig, Nicolás García Trillos, and Dejan Slepčev

Abstract In this work, we build a unifying framework to interpolate between density-driven and geometry-based algorithms for data clustering and, specifically, to connect the mean shift algorithm with spectral clustering at discrete and continuum levels. We seek this connection through the introduction of Fokker–Planck equations on data graphs. Besides introducing new forms of mean shift algorithms on graphs, we provide new theoretical insights on the behavior of the family of diffusion maps in the large sample limit as well as provide new connections between diffusion maps and mean shift dynamics on a fixed graph. Several numerical examples illustrate our theoretical findings and highlight the benefits of interpolating density-driven and geometry-based clustering algorithms.

1 Introduction

In this work we establish new connections between two popular but seemingly unrelated families of methodologies used in unsupervised learning. The first family that we consider is density based and includes mode seeking clustering approaches such as the mean shift algorithm introduced in [15] and reviewed in [9], while the second family is based on spectral geometric ideas applied to graph settings and includes methodologies such as Laplacian eigenmaps [3] and spectral clustering [32]. After discussing the mean shift algorithm in Euclidean space and reviewing a family of spectral methods for clustering on graphs, we seek these connections

K. Craig
University of California, Santa Barbara, CA, USA
e-mail: kcraig@math.ucsb.edu

N. García Trillos
University of Wisconsin, Madison, WI, USA
e-mail: garciatrillo@wisc.edu

D. Slepčev (✉)
Carnegie Mellon University, Pittsburg, PA, USA
e-mail: slepcev@math.cmu.edu

N. Bellomo et al. (eds.), *Active Particles, Volume 3*, Modeling and Simulation in Science, Engineering and Technology, https://doi.org/10.1007/978-3-030-93302-9_4

in two different ways. First, motivated by some heuristics at the continuum level (i.e., infinite data setting level), we take a suitable dynamic perspective and introduce appropriate interpolating *Fokker–Planck* equations on data graphs. This construction is inspired by the variational formulation in the Wasserstein space of Fokker–Planck equations at the continuum level and utilizes recent geometric formulations of PDEs on graphs. Second, we revisit the diffusion maps from [12] (with an extended range for the parameter indexing the family) and in particular show that, when parameterized conveniently and in the large data limit, the family of diffusion maps is closely related to the same family of continuum dynamics motivating our Fokker–Planck interpolations on graphs. At the finite data level, we show that by taking an extreme value of the parameter indexing the diffusion maps we can retrieve a specific graph version of the mean shift algorithm introduced in [25]. Our new theoretical insights are accompanied by extensive numerical examples aimed at illustrating the benefits of interpolating density and geometry-driven clustering algorithms.

To begin our discussion, let us recall that unsupervised learning is one of the fundamental settings in machine learning where the goal is to find structure in a data set $\mathcal{X}$ without the aid of any labels associated with the data. For example, if the data set $\mathcal{X}$ consisted of images of animals, a standard task in unsupervised learning would be to recognize the structure of groups in $\mathcal{X}$ without using information of the actual classes that may be represented in the data set (e.g., {dog, olinguito, caterpillar, …}); in the literature, this task is known as *data clustering* and will be our main focus in this chapter. Other unsupervised learning tasks include dimensionality reduction [41] and anomaly detection [22], among others.

When clusters are geometrically simple, for example, when they are dense sets of points separated in space, elementary clustering methods, like k-means or k-median clustering, are sufficient to identify the clusters. However, in practice, clusters are often geometrically complex due to the natural variations that objects belonging to the cluster may have and also due to invariances that some object classes posses. To handle such data sets, there is a large class of clustering algorithms, including the ones that will be explored in this chapter, which are described as two-step procedures consisting of an embedding step and an actual clustering step where a more standard, typically simple, clustering method is used on the embedded data. In mathematical terms, in the first step, the goal is to construct a map

$$\Psi : \mathcal{X} \to \mathcal{Y}$$

between the data points in $\mathcal{X}$ and a space $\mathcal{Y}$ (e.g., $\mathcal{Y} = \mathbb{R}^k$ for some small k, or in general a metric space) to "disentangle" the original data as much as possible, and in the second step, the actual clustering is obtained by running a simple clustering algorithm such as K-means (if the set $\mathcal{Y}$ is the Euclidean space, for example). While both steps are important and need careful consideration, it is in the first step, and specifically in the choice of Ψ, that most clustering algorithms differ from each other. For example, as we will see in Sect. 1.1, in some version of mean shift, Ψ is induced by gradient ascent dynamics of a density estimator starting at the different data points (when the data points are assumed to lie in Euclidean space), whereas in spectral methods for clustering in graph-based settings the map Ψ is typically built using the low-lying spectrum of a suitable graph Laplacian. At its heart, different choices of Ψ capture different heuristic interpretations of the loosely defined notion of "data cluster." Some constructions have a density-driven flavor (e.g., mean shift), while others are inspired by geometric considerations (e.g., spectral clustering). The notions of density-based and geometry-based algorithms, however, are not monolithic, and each individual algorithm has nuances and drawbacks that are important to take into account when deciding whether to use it or not in a given situation; in our numerical experiments, we will provide a series of examples that highlight some of the qualitative weaknesses of different clustering algorithms, density or geometry driven. Our main aim is to introduce a new mathematical framework to interpolate between these two seemingly unrelated families of clustering algorithms and to provide new insights for existing interpolations like *diffusion maps*.

In the rest of this introduction, we present some background material that is used in the remainder.

1.1 Mean Shift-Based Methods

Let $\mathcal{X} = \{x_1, \dots, x_n\}$ be a data set in $\mathbb{R}^d$. One heuristic way to define data "clusters" is to describe them as regions in space of high concentration of points separated by areas of low density. One algorithm that uses this heuristic definition is the *mean shift algorithm*; see [9]. In essence, mean shift is a hill climbing scheme that seeks the local modes of a density estimator constructed from the observed data in an attempt to identify data clusters.

To make the discussion more precise, let us first describe the setting where the points $x_1, \dots, x_n$ are obtained by sampling a probability distribution supported on the whole $\mathbb{R}^d$ with (unknown) smooth enough density $\rho : \mathbb{R}^d \to \mathbb{R}$. The first step in mean shift is to build an estimator for ρ of the form:

$$\hat{\rho}(x) := \frac{1}{n\delta^d} \sum_{i=1}^{n} \kappa\left(\frac{|x - x_i|}{\delta}\right),$$

where $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ is an appropriately normalized kernel, and $\delta > 0$ is a suitable bandwidth; for simplicity, we can take the standard Gaussian kernel $\kappa(s) = \frac{1}{\sqrt{2\pi}} \exp^{-s^2/2}$. Then, for every point x_i, one considers the iterations:

$$x_i(t+1) = x_i(t) + \nabla \log \hat{\rho}(x_i(t)) \tag{1.1}$$

starting at time $t = 0$ with $x_i(0) = x_i$. Each data point x_i is then mapped to its associated $x_i(T)$ for some user-specified T, implicitly defining in this way a map Ψ as described in the introduction. It is important to notice that mean shift is, in the way introduced above, a *monotonic* scheme, i.e., $\hat{\rho}(x_i(t))$ is non-decreasing as a function of $t \in \mathbb{N}$. In [8], this is shown to be a consequence of a deeper property that in particular relates mean shift with the expectation-maximization algorithm applied to a closely related problem. The name mean shift originates from the fact that iterate $x_i(t+1)$ in (1.1) coincides with the mean of some distribution that is centered around the iterate $x_i(t)$, and thus (1.1) can be described as a "mean shifting" scheme.

We now introduce the continuum analogue of the mean shift algorithm (1.1). Namely, (1.1) can be seen as the iterates of the Euler scheme, for time step $h = 1$ of the following ODE system:

$$\begin{cases} \dot{x}_i(t) = \nabla \log \hat{\rho}(x_i(t)), \quad t > 0 \\ x_i(0) = x_i. \end{cases} \tag{1.2}$$

In the remainder, we will abuse the terminology slightly and refer to the above continuous time dynamics as *mean shift*. We note that the $\hat{\rho}$ is monotonically increasing along the trajectories of (1.2).

To utilize the mean shift dynamics for clustering, for some prespecified time $T > 0$ (that at least theoretically can be taken to be infinity under mild conditions on ρ), we consider the embedding map:

$$\Psi_{MS}(x_i) := x_i(T), \quad x_i \in \mathcal{X}.$$

When the number of data points n in $\mathcal{X}$ is large, and the bandwidth h is small enough (but not too small), one can heuristically expect that the gradient lines of the density estimator $\hat{\rho}$ resemble those of the true density ρ; see for example [2]. In particular, with an appropriate tuning of bandwidth δ as a function of n, and with a large value of T defining the time horizon for the dynamics, one can expect Ψ_{MS} to send the original data points to a set of points that are close to the local modes of the density ρ. In short, mean shift is expected to cluster the original data set by assigning points to the same cluster if they belong to the same basin of attraction of the gradient ascent dynamics for the density ρ.

If the density ρ is supported on a manifold, $\mathcal{M}$, embedded in $\mathbb{R}^d$, and that information is available, one can consider mean shift dynamics restricted to the manifold. Indeed, to define manifold mean shift, we just need to consider the flow ODE (1.2), where ∇ is replaced by the gradient on $\mathcal{M}$, which for manifolds in $\mathbb{R}^d$ is just the projection of ∇ to the tangent space $T_x\mathcal{M}$. We notice that this extends to manifolds with boundary where at boundary point $x \in \partial\mathcal{M}$ one is projecting to the interior half-space $T_x^{in}\mathcal{M}$. We denote the projection to the tangent space (interior tangent space at the boundary) by $P_{T\mathcal{M}}$ and write the resulting ODE as

$$\begin{cases} \dot{x}_i(t) = P_{T\mathcal{M}} \nabla \log(\hat{\rho})(x_i(t)), \quad t > 0 \\ x_i(0) = x_i \in \mathcal{M}. \end{cases} \tag{1.3}$$

1.1.1 Lifting the Dynamics to the Wasserstein Space

Looking forward to our discussion in subsequent sections where we introduce mean shift algorithms on graphs, it is convenient to rewrite (1.2) in an alternative way using dynamics in the *Wasserstein* space $\mathcal{P}_2(\mathbb{R}^d)$. As it turns out, the ODE (1.2) is closely related to an ODE in $\mathcal{P}_2(\mathbb{R}^d)$ (i.e., the space of Borel probability measures over $\mathbb{R}^d$ with finite second moments). Precisely, we consider

$$\partial_t \mu_t + \operatorname{div}(\nabla \log(\hat{\rho})\mu_t) = 0, \quad t > 0, \tag{1.4}$$

with initial datum $\mu_0 = \delta_{x_i}$; equation (1.4) must be interpreted in the weak sense (see Chapter 8.1. in [1]). Indeed, when $\mu_0 = \delta_{x_i}$, it is straightforward to see that the solution to (1.4) is given by $\mu_t = \delta_{x_i(t)}$, where $x_i(\cdot)$ solves the ODE (1.2) in the base space $\mathbb{R}^d$. What is more, in the same way that (1.2) can be understood as the gradient descent dynamics for $-\log(\hat{\rho})$ in the base space $\mathbb{R}^d$, it is possible to interpret (1.4) directly as the gradient flow of the potential energy:

$$\mathcal{E}(\mu) := -\int_{\mathbb{R}^d} \log(\hat{\rho}(x))d\mu(x), \quad \mu \in \mathcal{P}_2(\mathbb{R}^d) \tag{1.5}$$

with respect to the Wasserstein metric d_W, which in dynamic form reads

$$d_W^2(\nu, \tilde{\nu}) = \inf_{t\in[0,1]\mapsto(\nu_t,\vec{V}_t)} \int_0^1 \int_{\mathbb{R}^d} |\vec{V}_t|^2 \, d\nu_t dt, \tag{1.6}$$

where the infimum is taken over all solutions $(\nu_t, \vec{V}_t)$ to the continuity equation

$$\partial_t \nu_t + \operatorname{div}(\nu_t \vec{V}_t) = 0,$$

with $\nu_0 = \nu$ and $\nu_1 = \tilde{\nu}$.

The previous discussion suggests the following alternative definition for the embedding map associated with mean shift:

$$\Psi_{MS}(x_i) := \mu_{i,T} \in \mathcal{P}_2(\mathbb{R}^d),$$

where in order to obtain $\mu_{i,T}$, we consider the gradient flow dynamics $\mathcal{E}$ in the Wasserstein space initialized at the point $\mu_0 = \delta_{x_i}$ (i.e., Eq. (1.4)). While this new interpretation may seem superfluous at first sight given that $\mu_{i,T} = \delta_{x_i(T)}$, we will later see that working in the space of probability measures is convenient, as this alternative representation motivates new versions of mean shift algorithms for data clustering on structures such as weighted graphs; see Sect. 2.2.

1.2 Spectral Methods

Let us now discuss another family of algorithms used in unsupervised learning that are based on ideas from spectral geometry. The input in these algorithms is a collection of edge weights w describing the similarities between data points in $\mathcal{X}$; we let $n = |\mathcal{X}|$. For simplicity, we assume that the weight function $w : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric and that all its entries are non-negative. We further assume that the weighted graph $\mathcal{G} = (\mathcal{X}, w)$ is connected in the sense that for every $x, x' \in \mathcal{X}$ there exists a path $x_0, \ldots, x_m \in \mathcal{X}$ with $x_0 = x$, $x_m = x'$ and $w(x_l, x_{l+1}) > 0$ for every $l = 0, \ldots, m-1$. At this stage, we do not assume any specific geometric structure in the data set $\mathcal{X}$ or on the weight function w (in Sect. 4, however, we focus our discussion on proximity graphs).

Let us now give the definition of well-known graph analogues of gradient, divergence, and Laplacian operators. To a function $\phi : \mathcal{X} \to \mathbb{R}$, we associate a *discrete gradient*, a function of the form $\nabla_{\mathcal{G}}\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$\nabla_{\mathcal{G}}\phi(x, x') := \phi(x') - \phi(x).$$

Given a function $U : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (i.e., a discrete vector field), we define its *discrete divergence* as the function $\mathrm{div}_{\mathcal{G}} U : \mathcal{X} \to \mathbb{R}$ given by

$$\mathrm{div}_{\mathcal{G}}\, U(x) := \frac{1}{2}\sum_{x'}(U(x', x) - U(x, x'))w(x, x').$$

With these definitions, we can now introduce the *unnormalized Laplacian* associated with the graph $\mathcal{G}$ as the operator $\Delta_{\mathcal{G}} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ defined according to

$$\Delta_{\mathcal{G}} := \mathrm{div}_{\mathcal{G}} \circ \nabla_{\mathcal{G}}$$

or more explicitly as

$$\Delta_{\mathcal{G}} u(x_i) = \sum_j (u(x_i) - u(x_j))w(x_i, x_j), \quad x_i \in \mathcal{X}, \quad u \in L^2(\mathcal{X}). \tag{1.7}$$

From the representation $\Delta_{\mathcal{G}} = \mathrm{div}_{\mathcal{G}} \circ \nabla_{\mathcal{G}}$, it is straightforward to verify that $\Delta_{\mathcal{G}}$ is a self-adjoint and positive semi-definite operator with respect to the $L^2(\mathcal{X})$ inner product (i.e., the Euclidean inner product in $\mathbb{R}^n$ after identifying real-valued functions on $\mathcal{X}$ with $\mathbb{R}^n$). It can also be shown that $\Delta_{\mathcal{G}}$ has zero as an eigenvalue with multiplicity equal to the number of connected components of $\mathcal{G}$ (in this case 1 by assumption); see [45]. Moreover, even when the multiplicity of the zero eigenvalue is uninformative about the group structure of the data set, the low-lying spectrum of $\Delta_{\mathcal{G}}$ still carries important geometric information for clustering. In particular, $\Delta_{\mathcal{G}}$'s small eigenvalues and their corresponding eigenvectors contain information on *bottlenecks* in $\mathcal{G}$ and on the corresponding regions that are separated by them; the connection between the spectrum of $\Delta_{\mathcal{G}}$ and the bottlenecks in $\mathcal{G}$ is expressed more precisely with the relationship between Cheeger constants and Fiedler eigenvalues; see [45]. With this motivation in mind, [3] introduced a nonlinear transformation of the data points known as *Laplacian eigenmap*:

$$x_i \in \mathcal{X} \longmapsto \begin{pmatrix} \phi_1(x_i) \\ \vdots \\ \phi_k(x_i) \end{pmatrix} \in \mathbb{R}^k,$$

where $\phi_1, \dots, \phi_k$ are the eigenvectors corresponding to the first k eigenvalues of $\Delta_{\mathcal{G}}$. The above Laplacian eigenmap and other similar transformations serve as the embedding map in the first step in most spectral methods for partitioning and data clustering. Said clustering algorithms have a rich history, and related ideas have been present in the literature for decades, see [32, 35, 45] and the references within. For example, some versions of spectral clustering consider a conformal transformation of the Laplacian eigenmap in which coordinates in the embedding space are rescaled differently according to corresponding eigenvalues and the choice of a timescale parameter. More precisely, one may consider

$$\hat{\Psi}_{SC}(x_i) := \begin{pmatrix} e^{-T\lambda_1}\phi_1(x_i) \\ \vdots \\ e^{-T\lambda_k}\phi_k(x_i) \end{pmatrix} \in \mathbb{R}^k, \quad x_i \in \mathcal{X},$$

for some $T > 0$, where in the above λ_l represents the eigenvalue corresponding to the eigenvector ϕ_l. In Sect. 2, we provide a dynamic interpretation of the map $\hat{\Psi}_{SC}$.

Remark 1.1 There are several other ways in the literature to construct the embedding maps Ψ from graph Laplacian eigenvectors. In [32], for example, an extra normalization step across eigenvectors is considered for each data point. By introducing this extra normalization step, one effectively maps the data points into the unit sphere in Euclidean space. The work [34] argues in favor of this type of normalization and proposes the use of an angular version of k-means clustering on the embedded data set. The work [17] also analyzes the geometric structure of spectral embeddings, both at the data level and at the continuum population level. The normalization step in [32] can also be motivated from a robustness to outliers perspective if one insists on running k-means with the ℓ^2 metric and not with for example the ℓ^1 metric. As discussed in the introduction, constructing a data embedding is only part of the full clustering problem. What metric and what clustering method should be used on the embedded data are important practical and theoretical questions. In subsequent sections, our embedded data points will have the form of probability vectors. Several metrics could then be used to cluster the embedded data points (TV, L^2, W_2, etc), each with its own advantages and disadvantages (theoretical or computational). The emphasis of our discussion in the rest of the chapter, however, will be on the embedding maps themselves. We leave the analysis of the effect of different metrics used to cluster the embedded data for future work.

1.2.1 Normalized Versions of the Graph Laplacian

Different versions of graph Laplacians can be constructed to include additional information about vertex degrees as well as to normalize the size of eigenvalues. We distinguish between two ways to normalize the graph Laplacian $\Delta_{\mathcal{G}}$. One is based on reweighing operators and the other on renormalizing edge weights.

Operator-Based Renormalizations To start, we first write the graph Laplacian $\Delta_{\mathcal{G}}$ in matrix form. For that purpose, let $W = [w(x_i, x_j)]_{i,j}$ be the matrix of weights, and let

$$d(x_i) = \sum_{x_j \neq x_i} w(x_i, x_j) \tag{1.8}$$

be the weighted degrees; in the remainder, we may also use the notation $d_i = d(x_i)$ whenever no confusion arises from doing so. Let $D = \operatorname{diag}(d_1, \dots, d_n)$ be the diagonal matrix of degrees. The Laplacian $\Delta_{\mathcal{G}}$ can then be written in matrix form as

$$\Delta_{\mathcal{G}} = D - W. \tag{1.9}$$

In terms of this matrix representation, the *normalized* graph Laplacian, as introduced in [45], can be written as

$$L = D^{-\frac{1}{2}} \Delta_{\mathcal{G}} D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}. \tag{1.10}$$

Notice that the matrix L is symmetric and positive semi-definite as it follows directly from the same properties for $\Delta_{\mathcal{G}}$. The *random-walk* graph Laplacian, on the other hand, is given by

$$L^{rw} = D^{-1} \Delta_{\mathcal{G}} = I - D^{-1} W. \tag{1.11}$$

Remark 1.2 It is straightforward to show that the matrix L^{rw} is similar to the matrix L, and thus the random-walk Laplacian has the same eigenvalues as the normalized graph Laplacian. Moreover, if we explicitly use the representation $L^{rw^T} = D^{1/2} L D^{-1/2}$, where L^{rw^T} is the transpose of L^{rw}, we can see that if $\tilde{u}$ is an eigenvector of L with eigenvalue λ, then $\phi = D^{1/2}\tilde{\phi}$ is an eigenvector of L^{rw^T} with eigenvalue λ. In particular, since L is symmetric, we can find a collection of vectors $\phi_1, \dots, \phi_n \in \mathbb{R}^n$ that form an orthonormal basis (with respect to the inner product $\langle D^{-1}\cdot, \cdot\rangle$) for $\mathbb{R}^n$ and where each of the ϕ_l is an eigenvector for L^{rw^T}.

We use the above observation in Sect. 1.2.2 and later at the beginning of Sect. 2.

Edge Weight Renormalizations Here, the idea is to adjust the weights of the graph $\mathcal{G} = (\mathcal{X}, w)$ and use one of the Laplacian normalizations introduced before on the new graph. One of the most popular families of graph Laplacian normalizations based on edge reweighing was introduced in [12] and includes the generators for the so-called *diffusion maps* which we now discuss.

For a given choice of parameter $\alpha \in (-\infty, 1]$, we construct new edge weights, $w_\alpha(x, y)$, as follows:

$$w_\alpha(x, y) := \frac{w(x, y)}{d(x)^\alpha d(y)^\alpha},$$

where recall d is the weighted vertex degree. On the new graph $(\mathcal{X}, w_\alpha)$, one can consider all forms of graph Laplacian discussed earlier. In the sequel, however, we follow [12] and restrict our attention to the reweighed random-walk Laplacian which in matrix form can be written as

$$L_\alpha^{rw} = I - D_\alpha^{-1} W_\alpha,$$

where W_α is the matrix of edge weights of the new graph and D_α its associated degree matrix, i.e., $d_\alpha(x) = \sum_{y \neq x} w_\alpha(x, y)$. We note that the weight matrix W_α is still symmetric.

Let Q_α^{rw} be the *weighted diffusion rate* matrix

$$Q_\alpha^{rw} := -C_\alpha L_\alpha^{rw},$$

where C_α is a positive constant that we introduce for modeling purposes. In particular, in Sect. 4, we will see that, in the context of proximity graphs on data sampled from a distribution on a manifold $\mathcal{M}$, by choosing the constant C_α appropriately, we can ensure a desirable behavior of Q_α^{rw} as the number of data points in $\mathcal{X}$ grows. Notice that we may alternatively write Q_α^{rw} as a function:

$$Q_\alpha^{rw}(x, y) := C_\alpha \begin{cases} \frac{w_\alpha(x,y)}{\sum_{z \neq x} w_\alpha(x,z)} & \text{if } x \neq y, \\ -1 & \text{if } x = y. \end{cases} \tag{1.12}$$

Remark 1.3 In this chapter, we consider the range $\alpha \in (-\infty, 1]$ and not $[0, 1]$ as usually done in the literature. One important point that we stress throughout the chapter is that by considering the interval $(-\infty, 1]$, we obtain an actual interpolation between density-based (in the form of some version of mean shift) and geometry-driven clustering algorithms.

1.2.2 More General Spectral Embeddings

Following Remark 1.2, for a given $\alpha \in (-\infty, 1]$, we consider an orthonormal basis (relative to the inner product $\langle D_\alpha^{-1}\cdot, \cdot\rangle$) $\phi_1, \ldots, \phi_n$ of eigenvectors of $L_\alpha^{rw^T}$ with corresponding eigenvalues $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$ and define

$$\hat{\Psi}'_\alpha(x_i) := \begin{pmatrix} e^{-T\lambda_1}\tilde{\phi}_1(x_i) \\ \vdots \\ e^{-T\lambda_k}\tilde{\phi}_k(x_i) \end{pmatrix} \in \mathbb{R}^k, \quad x_i \in \mathcal{X}, \tag{1.13}$$

where, in the above, $\tilde{\phi}_l = D_\alpha^{-1/2}\phi_l$. We can see that the map $\hat{\Psi}'_\alpha$ has the same form as the map $\hat{\Psi}_{SC}$ at the beginning of Sect. 1.2. In Sect. 2, we provide a more dynamic interpretation of the map $\hat{\Psi}'_\alpha$.

1.3 *Outline*

Having discussed the mean shift algorithm in Euclidean setting (or in general on a submanifold $\mathcal{M}$ of $\mathbb{R}^d$), as well as some spectral methods for clustering in the graph setting, in what follows we attempt to build bridges between geometry-based and density-driven clustering algorithms. Our first step is to introduce general data embedding maps Ψ associated with the dynamics induced by arbitrary rate matrices on $\mathcal{X}$. We will then define data graph analogues of mean shift dynamics. We do this in Sect. 2 where we define a new version of mean shift on graphs inspired by the discussion in Sect. 1.1.1 and review other versions of mean shift on graphs such as Quickshift [44] and KNF [25]. In Sect. 3.1, we discuss two versions of Fokker–Planck equations on graphs which serve as interpolating dynamics between geometry- and density-driven dynamics for clustering on data graphs. One version is based on a direct interpolation between diffusion and mean shift (the latter one as defined in Sect. 2.2.1) and is inspired by Fokker–Planck equations at the continuum level. The second version is an extended version of the diffusion maps from [12] obtained by appropriate reweighing and normalization of the data graph. In Sect. 3.2, we show that the KNF mean shift dynamics can be seen as a particular case of the family of diffusion maps when the parameter indexing this family is sent to negative infinity. This result is our first concrete connection between mean shift algorithms and spectral methods for clustering. In Sect. 4, we study the continuum limits of the Fokker–Planck equations introduced in Sect. 3 when the graph of interest is a proximity graph. This analysis will allow us to provide further insights into diffusion maps, mean shift, and spectral clustering. In Sect. 5, we present a series of numerical experiments aimed at illustrating some of our theoretical insights.

2 Mean Shift and Fokker–Planck Dynamics on Graphs

Consider a weighted graph $\mathcal{G} = (\mathcal{X}, w)$ as in Sect. (1.2). Let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$ which we identify with n-dimensional vectors. All of the dynamics we consider can be written as (continuous time) Markov chains on graphs.

Definition 2.1 A *continuous time Markov chain* $u : [0, T] \to \mathcal{P}(\mathcal{X})$ is a solution to the ordinary differential equation

$$\begin{cases} \partial_t u_t(y) = \sum_{x\in\mathcal{X}} u_t(x) Q(x, y), \quad t > 0 \\ u_0 = u^0, \end{cases} \tag{2.1}$$

where $u^0 \in \mathcal{P}(\mathcal{X})$ and $Q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *transition rate matrix* Q; that is, Q satisfies $Q(x, y) \geqslant 0$ for $y \neq x$ and $Q(x, x) = -\sum_{y\neq x} Q(x, y)$.

Notice that in terms of matrix exponentials, the solution to (2.1) can be written as

$$u_t(y) = \sum_{x\in\mathcal{X}} u^0(x)(e^{tQ})(x, y).$$

Remark 2.2 The operation on the right-hand side of the first equation in (2.1) can be interpreted as a matrix multiplication of the form $u_t Q$, where u_t is interpreted as a row vector. Alternatively, we can use the transpose of Q and write $Q^T u_t$ if we interpret u_t as a column vector.

Remark 2.3 (Conservation of Mass and Positivity) For any transition rate matrix, we have

$$\sum_{x\in\mathcal{X}} Q(x, y) = 0, \quad \forall y \in \mathcal{X},$$

which ensures that $\sum_x u_t(x) = \sum_x u_0(x) = 1$ for all $t \geqslant 0$. Note that, in practice, this is often accomplished by specifying the *off-diagonal* entries of the transition rate matrix, $Q(x, y)$ for $x \neq y$, and then setting the diagonal entries to equal opposite the associated diagonal degree matrix, $d(x, x) = \sum_{y\neq x} Q(x, y)$. Likewise, for any transition rate matrix, the fact that $Q(x, y) \geqslant 0$ for $y \neq x$ ensures that if $u_0(x) \geqslant 0$, then $u_t(x) \geqslant 0$ for all $t \geqslant 0$.

Remark 2.4 Notice that $Q = -\Delta_{\mathcal{G}}$, where we recall $\Delta_{\mathcal{G}}$ is defined in (1.7), is indeed a rate matrix and thus induces evolution equations in $\mathcal{P}(\mathcal{X})$. Likewise, the weighted diffusion rate matrix Q^{rw}_{α} from (1.12) is a rate matrix as introduced in Definition 2.1.

For a general rate matrix Q, we define the data embedding map:

$$\hat{\Psi}_Q(x_i) := u_{i,T,Q} \in \mathcal{P}(X), \quad x_i \in \mathcal{X}, \tag{2.2}$$

where $u_{i,T,Q}$ represents the solution of (2.1) when the initial condition $u^0 \in \mathcal{P}(\mathcal{X})$ is defined by $u^0(x) = 1$ for $x = x_i$ and $u^0(x) = 0$ otherwise. In the sequel, and specifically in our numerics section, we will use the $L^2(\mathcal{X})$ metric between the points $\hat{\Psi}_Q(x_i)$ in order to build clusters through K-means regardless of the rate matrix Q that we use to construct the embedding $\hat{\Psi}_Q$. Remember that by $L^2(\mathcal{X})$ distance we mean the quantity:

$$\sum_{x'} (u_{i,T,Q}(x') - u_{j,T,Q}(x'))^2.$$

In the next subsections, we discuss some specifics of the choice $Q = Q_\alpha^{rw}$ and then introduce two classes of rate matrices Q that give meaning to the idea of mean shift on graphs.

2.1 Dynamic Interpretation of Spectral Embeddings

When we take $Q = Q_\alpha^{rw}$, we abuse notation slightly and write $\hat{\Psi}_\alpha$ instead of $\hat{\Psi}_{Q_\alpha^{rw}}$ and $u_{i,T,\alpha}$ instead of $u_{i,T,Q_\alpha^{rw}}$. In the next proposition, we make the connection between the embedding map $\hat{\Psi}_\alpha$ and the spectral embedding $\hat{\Psi}'_\alpha$ from (1.13) explicit.

Proposition 2.5 *For every $\alpha \in (-\infty, 1]$, we can write*

$$u_{i,T,\alpha}(x) = \sum_{l=1}^{n} e^{-T\lambda_l} \frac{\phi_l(x_i)}{(d_\alpha(x_i))^{1/2}} \phi_l(x), \quad \forall x \in \mathcal{X},$$

where the $\phi_1, \ldots, \phi_n$ form an orthonormal basis for $\mathbb{R}^n$ (with respect to the inner product $\langle D_\alpha^{-1} \cdot, \cdot \rangle$) and each ϕ_l is an eigenvector of ${L_\alpha^{rw}}^T$ with eigenvalue λ_l. In other words, the coordinates of the vector $\hat{\Psi}'_\alpha(x_i)$ correspond to the representation of $\hat{\Psi}_\alpha(x_i)$ in the basis $\phi_1, \ldots, \phi_n$.

Proof This follows from a simple application of the spectral theorem using Remark 1.2. □

Remark 2.6 An alternative way to compare the points $\hat{\Psi}_\alpha(x_i)$ and $\hat{\Psi}_\alpha(x_j)$ is to compute their weighted distance:

$$\sum_{x'} \left(\frac{u_{i,T,Q}(x')}{\sqrt{d_\alpha(x')}} - \frac{u_{j,T,Q}(x')}{\sqrt{d_\alpha(x')}} \right)^2.$$

This construction is introduced in [12] and is referred to as *diffusion distance*. It is worth mentioning that, as pointed out in [12], the diffusion distance between $\hat{\Psi}_\alpha(x_i)$ and $\hat{\Psi}_\alpha(x_j)$ coincides with the Euclidean distance between $\hat{\Psi}'_\alpha(x_i)$ and $\hat{\Psi}'_\alpha(x_j)$ when $k = n$. Notice that the diffusion metric is conformal to the $L^2(\mathcal{X})$ metric.

Remark 2.7 We remark that the embedding maps $\hat{\Psi}_{SC}$ and $\hat{\Psi}_Q$ for $Q = -\Delta_{\mathcal{G}}$ are connected in a similar way as the maps $\hat{\Psi}'_\alpha$ and $\hat{\Psi}_\alpha$ are. Indeed, if we let $\phi_1, \ldots, \phi_n$ be an orthonormal basis for $L^2(\mathcal{X})$ consisting of eigenvectors of $\Delta_{\mathcal{G}}$ (remember that $\Delta_{\mathcal{G}}$ is positive semi-definite with respect to $L^2(\mathcal{X})$), then the coordinates of $\hat{\Psi}_{SC}(x_i)$ are precisely the coordinates of the representation of $u_{i,T,Q}$ in the basis $\phi_1, \ldots, \phi_n$.

2.2 *The Mean Shift Algorithm on Graphs*

In the next two subsections, we discuss two different ways to introduce mean shift on $\mathcal{G} = (\mathcal{X}, w)$. In both cases, we define an associated rate matrix Q.

2.2.1 Mean Shift on Graphs as Inspired by Wasserstein Gradient Flows

The discussion in Sect. 1.1.1 shows that the mean shift dynamics can be viewed as a gradient flow in the spaces of probability measures endowed with Wasserstein metric. Recent works [14, 29] provide a way to consider Wasserstein type gradient flows which are restricted to graphs. This allows one to take advantage of the information about the geometry of data that their initial distribution provides. More importantly, for our considerations, it allows one to combine the mean shift and spectral methods.

The notion of Wasserstein metric on graphs introduced by Maas [29] provides the desired framework. Here, we will consider the upwind variant of the Wasserstein geometry on graphs introduced in [14] since it avoids the problems that the metric of [29] has when dealing with the continuity equations on graphs, see Remark 1.2 in [14].

In particular, we actually consider a *quasi-metric* on $\mathcal{P}(\mathcal{X})$ (and not a metric) defined by

$$\hat{d}^2_W(v, \tilde{v}) := \inf_{t\in[0,1]\mapsto(v_t, V_t)} \int_0^1 \sum_{x,y} |V_t(x, y)_+|^2 \, \overline{v}_t(x, y) w(x, y) dt,$$

where the infimum is taken over all solutions (v_t, V_t) to the discrete continuity equation:

$$\partial_t v_t + \mathrm{div}_{\mathcal{G}}(\overline{v}_t \cdot V_t) = 0,$$

with $v_0 = v$ and $v_1 = \tilde{v}$ and where V_t is anti-symmetric for all t. In the above, the constant $C_{ms} > 0$ is introduced for modeling purposes and will become relevant in Sect. 4. We use the upwinding interpolation [11, 14]:

$$\overline{v}_t(x, y) := \begin{cases} v_t(x) & \text{if } V_t(x, y) \geqslant 0, \\ v_t(y) & \text{if } V_t(x, x) < 0 \end{cases} \tag{2.3}$$

and interpret the discrete vector field $\overline{v}_t \cdot V_t$ as the elementwise product of $\overline{v}_t$ and V_t. Finally, we use a_+ to denote the positive part of the number a.

Next, we consider the general potential energy:

$$\hat{\mathcal{E}}(u) := -\sum_{x \in \mathcal{X}} B(x)u(x), \quad u \in \mathcal{P}(\mathcal{X}),$$

for some $B : \mathcal{X} \to \mathbb{R}$. This energy serves as an analogue of (1.5).

Following the analysis and geometric interpretation in [14], it is possible to show that the gradient flow of $\hat{\mathcal{E}}$ with respect to the quasi-metric $\hat{d}_W$ takes the form:

$$\partial_t u_t(y) = \sum_{x \in \mathcal{X}} u_t(x) Q^B(x, y), \tag{2.4}$$

where Q^B is the rate matrix defined by

$$Q^B(x, y) := \begin{cases} (B(y) - B(x))_+ w(x, y), & \text{for } x \neq y, \\ -\sum_{z \neq x} (B(z) - B(x))_+ w(x, z), & \text{for } x = y. \end{cases} \tag{2.5}$$

In Sect. 4.1, we explore the connection between the graph mean shift dynamics (2.4) and the mean shift dynamics on an m-dimensional submanifold of $\mathbb{R}^d$ as introduced in Sect. 1.4. This is done in the context of proximity graphs over a data set $\mathcal{X} = \{x_1, \ldots, x_n\}$ obtained by sampling a distribution with density ρ on $\mathcal{M}$. In particular, we formally show that the graph mean shift dynamics converges to the continuum one if

$$B(x) = -\frac{C_{ms}}{\rho(x)}, \quad x \in \mathcal{M}.$$

In practice, however, since ρ is in general unavailable, ρ above can be replaced with a density estimator $\hat{\rho}$. Given such an estimator $\hat{\rho}$ (which in principle can be considered on general graphs, not just ones embedded in $\mathbb{R}^d$), we define the transition kernel for the *graph mean shift* dynamics as

$$Q^{ms}(x, y) := C_{ms} \begin{cases} \left(-\frac{1}{\hat{\rho}(y)} + \frac{1}{\hat{\rho}(x)}\right)_+ w(x, y), & \text{for } x \neq y, \\ -\sum_{z \neq x} \left(-\frac{1}{\hat{\rho}(z)} + \frac{1}{\hat{\rho}(x)}\right)_+ w(x, z), & \text{for } x = y, \end{cases} \tag{2.6}$$

where the constant $C_{ms} > 0$ just sets the timescale. It will be specified in Sect. 4.1 so that the equation has the desired limit as $n \to \infty$. For data in $\mathbb{R}^d$, it is natural to consider a kernel density estimator $\hat{\rho}$. In particular, in all of our experiments, we consider the following kernel density estimator:

$$\hat{\rho}_\delta(x) := \frac{1}{n} \sum_{y \in \mathcal{X}} \psi_\delta(x - y), \quad \psi_\delta(x) = \frac{1}{(2\pi)^{m/2}\delta^m} e^{-|x|^2/(2\delta^2)}. \tag{2.7}$$

For an abstract graph $\mathcal{G}$ (i.e., $\mathcal{X}$ is not necessarily a subset of Euclidean space), one can consider the degree of the graph at each $x \in \mathcal{X}$ as a substitute for $\hat{\rho}$ in the above expressions.

From the previous discussion and in direct analogy with the discussion in Sect. 1.1.1, we introduce the data embedding map:

$$\hat{\Psi}_{ms}(x_i) := u_{i,T,Q^{ms}} \in \mathcal{P}(\mathcal{X}). \tag{2.8}$$

Remark 2.8 The quasi-metric $\hat{d}_W$ and the geometry of the PDEs on graphs that it induces have been studied in [14]. One important point made in that paper (see Remark 1.2. in [14]) is that the support of the solution to the induced equations may change and move as time increases. For us, this property is essential as we initialize our graph mean shift dynamics at Diracs located at each of the data points.

We now provide a couple of illustrations comparing the mean shift dynamics (1.2) and the graph mean shift dynamics defined by (2.6). We consider data on manifold $\mathcal{M} = [0, 4] \times \{0, 0.7\}$. The measure ρ has uniform density on the two line segments. We consider 280 data points sampled from ρ, Fig. 1, and sampled from ρ with Gaussian noise of variance 0.1 in vertical direction, Fig. 3a.

We compare the dynamics on $\mathcal{M}$ for different bandwidths δ of the kernel density estimator. In particular, we consider a value of δ that is small enough for the strips to be seen as separate and a value of delta that is large enough for the strips to be considered together, Fig. 2. For large δ, we see rather different behavior of the two dynamics. The standard mean shift quickly mixes the data from the two lines and the information about the two clusters is lost. On the other hand, while the driving

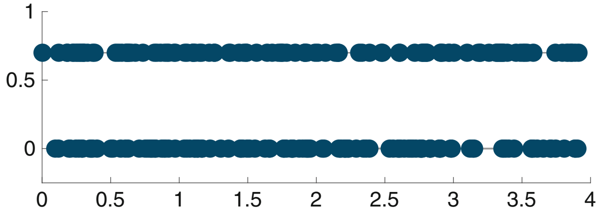
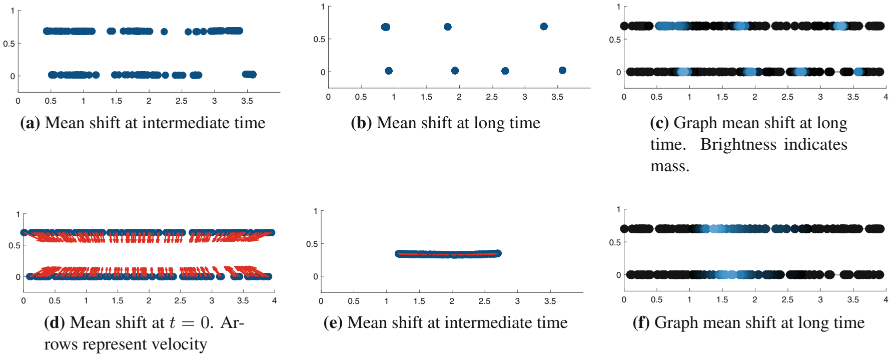


Fig. 1 Initial data for the experiments below. There are 280 points sampled from a uniform distribution on two line segments



(a) Mean shift at intermediate time

(b) Mean shift at long time

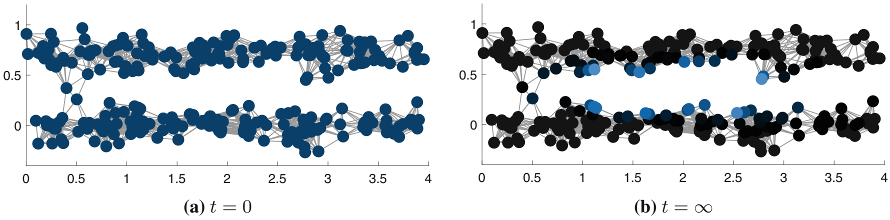
(c) Graph mean shift at long time. Brightness indicates mass.

(d) Mean shift at $t = 0$. Arrows represent velocity

(e) Mean shift at intermediate time

(f) Graph mean shift at long time

Fig. 2 We compare the dynamics for the mean shift (1.2) and graph mean shift (2.4)–(2.6). The top row shows the dynamics for $\delta = \frac{1}{4}$ bandwidth of the KDE. Both approaches give similar results. The stripes evolve independently and there are spurious local maxima due to randomness. The bottom row shows the dynamics for a larger $\delta = \frac{1}{\sqrt{2}}$. The KDE has a unique maximum. Mean shift quickly mixes the stripes into one, which then collapses to a point. On the other hand, since graph mean shift dynamics is constrained to the sample points the stripes do not mix and a single mode is identified in each stripe. (**a**) Mean shift at intermediate time . (**b**) Mean shift at long time. (**c**) Graph mean shift at long time. Brightness indicates mass. (**d**) Mean shift at $t = 0$. Arrows represent velocity. (**e**) Mean shift at intermediate time. (**f**) Graph mean shift at long time



(a) $t = 0$

(b) $t = \infty$

Fig. 3 Graph mean shift for $\delta = \frac{1}{\sqrt{2}}$. If noise is added to the data above, most of the dynamics behave as before. The exception is shown. The graph mean shift does not reach the modes as on Fig. 2f. Namely due to geometric roughness of the data the dynamics gets trapped at blue points. (**a**) $t = 0$. (**b**) $t = \infty$

force is the same, in the graph mean shift the dynamics is restricted to the data, thus preventing the mixing. In particular, separate modes are identified in each clump.

We note that this desirable behavior is somewhat fragile when noise is present, Fig. 3b. In particular, the roughness of the boundary prevents the mass to reach the mode. We will discuss later that this is mitigated by adding a bit of diffusion to the dynamics, see Sect. 5.2.6.

2.2.2 Quickshift and KNF

There are some alternative definitions of mean shift on graphs that are popular in the literature. One such algorithm is Quickshift [44], which is similar to an earlier algorithm by Koontz, Narendra, and Fukunaga [25]. Both algorithms can be described as hill climbing iterative algorithms for the maximization of a potential function $\hat{B}$.

Let $\hat{B} : \mathcal{X} \to \mathbb{R}$ be the potential for which we want to define "gradient ascent dynamics" along the graph $(\mathcal{X}, w)$. Let $\hat{D}(x, y) \geqslant 0$ be a notion of "distance" between points x and y which is typically defined through the weights w. Both the Quickshift and KNF algorithms have a Markov chain interpretation that we describe in a general form that allows for the (unlikely) existence of non-unique maximizers of $\hat{B} : \mathcal{X} \to \mathbb{R}$ around a given node $x \in \mathcal{X}$. To describe the associated rate matrices, let us define for every $x \in \mathcal{X}$ the sets

$$M_{QS,x} := \left\{ y \in \mathcal{X} \ : \ y \text{ maximizes: } \frac{1}{\hat{D}(x, y)} \mathbf{1}_{\hat{B}(y) > \hat{B}(x)} \right\}$$

and

$$M_{KNF,x} := \left\{ y \in \mathcal{X} \ : \ y \text{ maximizes: } \frac{(\hat{B}(y) - \hat{B}(x))_+}{\hat{D}(x, y)} \mathbf{1}_{\hat{D}(x,y) < r} \right\}.$$

The Quickshift and KNF algorithms are then the paths in the Markov chains with rate matrices:

$$Q_{QS}(x, y) = \begin{cases} \frac{1}{\sharp M_{QS,x}} & \text{if } y \in M_{QS,x}, \\ -1 & \text{if } y = x, \\ 0 & \text{otherwise}, \end{cases} \tag{2.9}$$

$$Q_{KNF}(x, y) = \begin{cases} \frac{1}{\sharp M_{KNF,x}} & \text{if } y \in M_{KNF,x}, \\ -1 & \text{if } y = x, \\ 0 & \text{otherwise}, \end{cases} \tag{2.10}$$

respectively. In Sect. 3.2, we establish a connection between the family of rate matrices Q_α^{rw} and Q_{KNF}.

3 Fokker–Planck Equations on Graphs

3.1 Fokker–Planck Equations on Graphs via Interpolation

The first type of interpolation between density- and geometry-driven clustering algorithms that we discuss in this chapter is based on a direct interpolation of the rate matrices Q^{ms} and Q_1^{rw}. Namely, for $\beta \in [0, 1]$, we consider

$$Q_\beta := \beta Q^{ms} + (1-\beta) Q_1^{rw}. \tag{3.1}$$

It is straightforward to see that the resulting Q_β continues to be a rate matrix and as such it induces dynamics in the space $\mathcal{P}(\mathcal{X})$. We can then use the framework from Sect. 2 and abuse notation slightly to write $\hat{\Psi}^\beta$ instead of $\hat{\Psi}_{Q_\beta}$ as well as $u_{i,T,\beta}$ instead of u_{i,T,Q_β}.

The choice of rate matrix Q_β is motivated by the Fokker–Planck equation:

$$\partial_t f_t = \beta \text{div}(\nabla\phi f_t) + (1-\beta)\Delta f_t$$

on a submanifold $\mathcal{M}$ of $\mathbb{R}^d$, which in the context of Sect. 4 can be proved to be a formal continuum limit of the evolution induced by Q_β as the number of data points grows. On the other hand, we notice that when we take $\beta = 1$ in Q_β, we recover the mean shift dynamics from Sect. 2.2.1. If on the contrary we set $\beta = 0$, we obtain the dynamics induced by the rate matrix Q_1^{rw}, which, at least in the context of Sect. 4, can be shown to be connected in the large sample limit to the heat equation on a manifold $\mathcal{M}$ where the data density plays no role.

3.2 Fokker–Planck Equation on Graphs via Reweighing and Connections to Graph Mean Shift

Another interpolation between density-driven and geometry-based clustering dynamics is induced by the family of rate matrices $\{Q_\alpha^{rw}\}_{\alpha\in(-\infty,1]}$. Indeed, in Sect. 4.2, we prove that in the proximity graph setting, the discrete dynamics associated with the rate matrices Q_α^{rw} are closely related, in the large data limit, to the same family of Fokker–Planck equations at the continuum level mentioned in Sect. 3.1. What is more, without taking a large sample limit, we see that the family $\{Q_\alpha^{rw}\}_{\alpha\in(-\infty,1]}$ interpolates between Q_1^{rw} and a rate matrix inducing graph mean shift dynamics, only that this time the version of mean shift that is meaningful is a particular case of the KNF formulation from Sect. 2.2.2. We prove this in the next proposition.

Proposition 3.1 *Let $(\mathcal{X}, w)$ be an arbitrary weighted graph satisfying the conditions at the beginning of Sect. 1.2. Set $C_\alpha = 1$ for every $\alpha \in (-\infty, 1]$. Then,*

$$\lim_{\alpha\to-\infty} Q^{rw}_{\alpha} = Q^{rw}_{-\infty}, \tag{3.2}$$

where

$$Q^{rw}_{-\infty}(x,y) := -\mathbf{1}_{y=x} + \begin{cases} \frac{w(x,y)}{\sum_{z\in M_{KNF,x}} w(x,z)} & \textit{if } y \in M_{KNF,x} \\ 0 & \textit{otherwise,} \end{cases} \tag{3.3}$$

where in the definition of $M_{KNF,x}$ we are using $\hat{B}(z) = d(z)$, $\hat{D}(x,y) = 1$ if $w(x,y) > 0$ and $\hat{D}(x,y) = \infty$ if $w(x,y) = 0$, and $r > 1$.

We notice that this is essentially the KNF rate matrix defined in (2.10) with only a difference in the way ties are broken when the maximum of d around a point is not unique. This distinction is mostly irrelevant since generically we may expect no ties. On the other hand, if for some reason there are ties but the non-zero weights in the graph are equal, then the two tie-breaking rules coincide.

Proof As the cases are analogous, let us consider only the case $y \neq x$. Note that

$$Q^{rw}_{\alpha}(x,y) = \frac{w_\alpha(x,y)}{\sum_{z\neq x} w_\alpha(x,z)} = \frac{w(x,y)d(x)^{-\alpha}d(y)^{-\alpha}}{\sum_{z\neq x} w(x,z)d(x)^{-\alpha}d(z)^{-\alpha}} = \frac{w(x,y)d(y)^{-\alpha}}{\sum_{z\neq x} w(x,z)d(z)^{-\alpha}}.$$

If $y \notin M_{KNF,x}$, consider $z \in M_{KNF,x}$. Then,

$$Q^{rw}_{\alpha}(x,y) \leqslant \frac{w(x,y)}{w(x,z)}\left(\frac{d(y)}{d(z)}\right)^{-\alpha} \to 0 \quad \text{as } \alpha \to -\infty.$$

If $y \in M_{KNF,x}$, then

$$Q^{rw}_{\alpha}(x,y) = \frac{w(x,y)}{\sum_{z\neq x} w(x,z)\left(\frac{d(z)}{d(y)}\right)^{-\alpha}} \to \frac{w(x,y)}{\sum_{z\in M_{KNF,x}} w(x,z)} \quad \text{as } \alpha \to -\infty.$$

□

4 Continuum Limits of Fokker–Planck Equations on Graphs and Implications

In this section, we further study the Fokker–Planck equations introduced in Sect. 3 and discuss their connection with Fokker–Planck equations at the continuum level. For such connection to be possible, we impose additional assumptions on the graph $\mathcal{G} = (\mathcal{X}, w)$. In particular, we assume that $\mathcal{G}$ is a *proximity graph* on $\mathcal{X} = \{x_1, \ldots, x_n\}$, where the x_i are assumed to be i.i.d. samples from a distribution on a smooth compact m-dimensional manifold without boundary $\mathcal{M}$ embedded in

$\mathbb{R}^d$ and having density $\rho : \mathcal{M} \to \mathbb{R}$ with respect to the volume form on $\mathcal{M}$. By proximity graph, we mean that the weights $w(x_i, x_j)$ are defined according to

$$w(x, y) := \eta_\varepsilon(|x - y|), \quad \eta_\varepsilon(r) = \frac{1}{\varepsilon^m}\eta\left(\frac{r}{\varepsilon}\right), \tag{4.1}$$

where $\varepsilon > 0$ is a bandwidth appropriately scaled with the number of samples n, η is a function $\eta : [0, \infty) \to [0, \infty)$ with compact support, and $|x - y|$ denotes the Euclidean distance between x and y.

4.1 Continuum Limit of Mean Shift Dynamics on Graphs

In order to formally derive the large sample limit of equation (2.4), we study the action of Q^{ms} on a smooth function $u : \mathcal{M} \to \mathbb{R}$. That is, we compute

$$\sum_{x\in\mathcal{X}} u(x)Q^{ms}(x, y) = -C_{ms} \sum_{x\in\mathcal{X}} \left[(B(x) - B(y))_+ u(y) - (B(x) - B(y))_- u(x)\right] w(x, y)$$

as $n \to \infty$ and $\varepsilon \to 0$ at a slow enough rate. Since our goal below is to deduce formal continuum limits, we will assume that $\mathcal{M}$ is flat. We note that when $\mathcal{M}$ is a smooth manifold, the deflection of the manifold from the tangent space is at most quadratic, and thus the error introduced is small when ε is small. In this way, we can avoid using the notation and constructions from differential geometry as well as some approximation arguments that obscure the reason why the limit holds. Providing a rigorous argument for the convergence of the dynamics remains an open problem.

In what follows, we use $\rho_n = \frac{1}{n}\sum_{x\in\mathcal{X}} \delta_x$ to denote the empirical distribution on the data points; here, we use the notation ρ_n to highlight the connection between the data points and the density function ρ. We also consider the constants

$$C_{ms} = \frac{1}{n\varepsilon^2\sigma_{\eta'}}, \quad \sigma_{\eta'} = \frac{1}{2m}\int_{\mathbb{R}^m} |z|^2\eta(|z|)dz,$$

and assume that the potential B is a $C^3(\mathcal{M})$ function. With the above definitions, we can explicitly write

$$-\sum_{x\in\mathcal{X}} u(x)Q^{ms}(x, y) = \frac{1}{n\varepsilon^{m+2}\sigma_{\eta'}} \sum_{x\in\mathcal{X}} \left[(B(x) - B(y))_+ u(y) - (B(x) - B(y))_- u(x)\right] \times \eta\left(\frac{|x - y|}{\varepsilon}\right) \tag{4.2}$$

and using the smoothness of u and B equate the above to

$$
\begin{aligned}
&= \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \left[(B(x) - B(y))_+ u(y) - (B(x) - B(y))_- u(x)\right] \eta\left(\frac{|x-y|}{\varepsilon}\right) \rho_n(x)dx \\
&= \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \big[(B(x) - B(y))_+ (u(y) - u(x)) \\
&\quad + \big((B(x) - B(y))_+ - (B(x) - B(y))_-\big) u(x)\big] \eta\left(\frac{|x-y|}{\varepsilon}\right) \rho_n(x)dx \\
&= \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \left[(B(x) - B(y))_+ (u(y) - u(x)) + (B(x) - B(y))\, u(x)\right] \eta \\
&\quad \times \left(\frac{|x-y|}{\varepsilon}\right) \rho_n(x)dx \\
&\approx \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \left[(B(x) - B(y))_+ \langle \nabla u(y), y - x\rangle + \langle \nabla B(y), x - y\rangle u(x)\right] \eta \\
&\times \left(\frac{|x-y|}{\varepsilon}\right) \rho_n(x)dx \\
&\quad + \frac{1}{2\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \Big[(B(x) - B(y))_+ \langle D^2 u(y)(x-y), y - x\rangle \\
&\quad + \langle D^2 B(y)(x-y), x - y\rangle u(x)\Big] \eta\left(\frac{|x-y|}{\varepsilon}\right) \rho_n(x)dx \\
&=: A_1 + A_2 + A_3 + A_4.
\end{aligned}
$$

Next, we analyze each of the terms A_1, A_2, A_3, and A_4. For A_1, we see that

$$
\begin{aligned}
A_1 &\approx \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} (B(x) - B(y))_+ \langle \nabla u(y), y - x\rangle \eta\left(\frac{|x-y|}{\varepsilon}\right) \rho(x)dx \\
&\approx -\frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\langle x-y, \nabla B(y)\rangle \geqslant 0} \langle \nabla u(y), x - y\rangle \langle x - y, \nabla B(y)\rangle \eta\left(\frac{|x-y|}{\varepsilon}\right) \rho(x)dx \\
&\approx -\frac{\rho(y)}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\langle x-y, \nabla B(y)\rangle \geqslant 0} \langle \nabla u(y), x - y\rangle \langle x - y, \nabla B(y)\rangle \eta\left(\frac{|x-y|}{\varepsilon}\right) dx \\
&= -\frac{\rho(y)}{\sigma_{\eta'}} \int_{\langle z, v\rangle \geqslant 0} \langle v', z\rangle \langle z, v\rangle \eta(|z|)dz,
\end{aligned}
\tag{4.3}
$$

where $v = \nabla B(y)$ and $v' = \nabla u(y)$; notice that in the first line we have replaced the empirical measure ρ_n with the measure $\rho(x)dx$ (introducing some estimation error) and in the second line we have considered a Taylor expansion of B around y. On the other hand, notice that

$$
\int_{\langle z, v\rangle \geqslant 0} \langle v', z\rangle \langle z, v\rangle \eta(|z|)dz = \langle Sv, v'\rangle,
$$

where S is a rank one symmetric matrix which can be written as $S = a\zeta \otimes \zeta$ for some vector ζ and some scalar a. Now, $a\langle \zeta, v\rangle^2$ is equal to

$$\begin{aligned}\langle Sv, v\rangle = \int_{\langle z,v\rangle \geqslant 0} \langle v, z\rangle^2 \eta(|z|)dz = |v|^2 \frac{1}{2m} \sum_{l=1}^{m} \int \langle e_l, z\rangle^2 \eta(|z|)dz \\ = |v|^2 \frac{1}{2m} \int |z|^2 \eta(|z|)dz = \sigma_{\eta'} |v|^2.\end{aligned}$$

The above computation shows that ζ can be taken to be v, and $a = \frac{\sigma_\eta'}{|v|^2}$. Thus,

$$A_1 \approx -\rho(y) \left\langle \nabla u(y), \nabla B(y) \right\rangle . \tag{4.4}$$

Regarding A_2, we have

$$A_2 \approx \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \langle \nabla B(y), x - y\rangle u(x) \eta \left(\frac{|x-y|}{\varepsilon} \right) \rho(x)dx,$$

introducing an estimation error to replace the integration with respect to the empirical measure with integration with respect to the measure $\rho(x)dx$. We can further decompose the computation introducing an approximation error:

$$A_2 \approx A_{21} + A_{22} + A_{23},$$

where

$$\begin{aligned}A_{21} &:= \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \langle \nabla B(y), x - y\rangle \langle \nabla u(y), x - y\rangle \eta \left(\frac{|x-y|}{\varepsilon} \right) \rho(y)dx, \\ A_{22} &:= \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \langle \nabla B(y), x - y\rangle u(y) \eta \left(\frac{|x-y|}{\varepsilon} \right) \rho(y)dx, \\ A_{23} &:= \frac{1}{\varepsilon^{m+2}\sigma_{\eta'}} \int_{\mathcal{M}} \langle \nabla B(y), x - y\rangle \langle \nabla \rho(y), x - y\rangle u(y) \eta \left(\frac{|x-y|}{\varepsilon} \right) dx.\end{aligned}$$

By symmetry, the term A_{22} is seen to be equal to zero. On the other hand, the terms A_{21} and A_{23} are computed similarly to the second expression in (4.3) only that in this case there is no sign constraint in the integral. From a simple change of variables, we can see that for arbitrary vectors v and v', we have

$$\int_{\langle z,v\rangle \geqslant 0} \langle v', z\rangle \langle z, v\rangle \eta(|z|)dz = \int_{\langle z,v\rangle \leqslant 0} \langle v', z\rangle \langle z, v\rangle \eta(|z|)dz.$$

In particular,

$$\int \langle v', z\rangle\langle z, v\rangle\eta(|z|)dz = 2\int_{\langle z,v\rangle\geqslant 0}\langle v', z\rangle\langle z, v\rangle\eta(|z|)dz = 2\sigma'_\eta\langle v, v'\rangle,$$

and thus

$$A_{21} = 2\rho(y)\langle\nabla B(y), \nabla u(y)\rangle, \quad A_{23} = 2u(y)\langle\nabla B(y), \nabla\rho(y)\rangle.$$

In summary,

$$A_2 \approx 2\rho(y)\langle\nabla B(y), \nabla u(y)\rangle + 2u(y)\langle\nabla B(y), \nabla\rho(y)\rangle. \tag{4.5}$$

It is straightforward to see that $A_3 = O(\varepsilon)$ and so for our computation we can treat A_3 as zero:

$$A_3 \approx 0. \tag{4.6}$$

For the final term A_4, we start by introducing an estimation error to write

$$A_4 \approx \frac{1}{2\varepsilon^{m+2}\sigma_{\eta'}}\int_{\mathcal{M}}\langle D^2B(y)(y-x), x-y\rangle u(x)\eta\left(\frac{|x-y|}{\varepsilon}\right)\rho(x)dx.$$

We can further replace the term $u(x)$ with $u(y)$ (and $\rho(x)$ with $\rho(y)$) in the formula above. This replacement introduces an $O(\varepsilon)$ term that we can ignore. It follows that

$$\begin{aligned} A_4 &\approx u(y)\rho(y)\frac{1}{2\varepsilon^{m+2}\sigma_{\eta'}}\int_{\mathcal{M}}\langle D^2B(y)(x-y), x-y\rangle\eta\left(\frac{|x-y|}{\varepsilon}\right)dx \\ &= u(y)\rho(y)\frac{1}{2\sigma_{\eta'}}\int\langle D^2B(y)z, z\rangle\eta(|z|)dz \\ &= u(y)\rho(y)\Delta B(y). \end{aligned}$$

Combining the above estimate with (4.4), (4.5), and (4.6), we see that

$$\begin{aligned} \sum_{x\in\mathcal{X}} u(x)Q^{ms}(x,y) &\approx -\left(\rho(y)\left\langle\nabla u(y), \nabla B(y)\right\rangle + 2u(y)\langle\nabla B(y), \nabla\rho(y)\rangle\right. \\ &\qquad \left. + u(y)\rho(y)\Delta B(y)\right) \\ &= -\frac{1}{\rho}\operatorname{div}\left(u\rho^2\nabla B\right). \end{aligned}$$

Note that the graph dynamics takes place on the provided data points, that is, on $\mathcal{P}(\mathcal{X}) \subset \mathcal{P}(\mathcal{M})$. As $n \to \infty$, $\mathcal{P}(\mathcal{X})$ approximates $\mathcal{P}(\mathcal{M})$. This partly explains that had we carried out the argument above in the full manifold setting the resulting dynamics would be restricted to the manifold and in particular both the divergence and the gradient above would take place on $\mathcal{M}$. That is, for data on a manifold,

$$\sum_{x\in\mathcal{X}} u(x)Q^{ms}(x,y) \approx -\frac{1}{\rho}\mathrm{div}_{\mathcal{M}}\left(u\rho^2\nabla_{\mathcal{M}}B\right).$$

Remark 4.1 Notice that with the choice $B(x) := \log(\rho(x))$, the above becomes

$$-\frac{1}{\rho}\mathrm{div}_{\mathcal{M}}\left(u\rho\nabla_{\mathcal{M}}\rho\right),$$

whereas with the choice $B(x) = -\frac{1}{\rho(x)}$, we get

$$-\frac{1}{\rho}\mathrm{div}_{\mathcal{M}}\left(u\rho\nabla_{\mathcal{M}}\log(\rho)\right).$$

The above analysis suggests that the formal continuum limit of the evolution (2.4) when $B = -\frac{1}{\rho}$ is the PDE:

$$\partial_t u_t = -\frac{1}{\rho}\mathrm{div}_{\mathcal{M}}(u_t\rho\nabla_{\mathcal{M}}\log(\rho)).$$

Notice, however, that the solution u_t of the above equation must be interpreted as a "density" with respect to the measure $\rho(x)dVol_{\mathcal{M}}$ ($\rho(x)dx$ in the flat case). Thus, in terms of "densities" with respect to $dVol_{\mathcal{M}}$, we obtain

$$\partial_t f_t = -\mathrm{div}_{\mathcal{M}}(f_t\nabla_{\mathcal{M}}\log(\rho)),$$

where $f_t := u_t\rho$. We recognize this latter equation as the PDE describing the mean shift dynamics (1.3).

4.2 Continuum Limits of Fokker–Planck Equations on Graphs

In this section, we formally derive the large sample limit of the two types of Fokker–Planck equations on $\mathcal{G}$ that we consider in this chapter, i.e., Eq. (2.1) when $Q = Q_\alpha^{rw}$ (for $\alpha \in (-\infty, 1]$) and when $Q = Q_\beta$ (for $\beta \in [0, 1]$).

We start our computations by pointing out that after appropriate scaling and under some regularity conditions on the density ρ, the diffusion operator L_α^{rw} converges toward the differential operator:

$$\mathcal{L}_\alpha v := -\frac{1}{\rho^{2(1-\alpha)}}\mathrm{div}_{\mathcal{M}}(\rho^{2(1-\alpha)}\nabla_{\mathcal{M}}v). \tag{4.7}$$

To be precise, if we set

$$C_\alpha = \frac{1}{\sigma_\eta \varepsilon^2}, \quad \sigma_\eta := \int_{\mathbb{R}^m} |z|^2 \eta(|z|) dz \Big/ \int_{\mathbb{R}^m} \eta(|z|) dz,$$

then, for all smooth $v : \mathcal{M} \to \mathbb{R}$, we have

$$-\sum_{y \in \mathcal{X}} Q_\alpha^{rw}(x, y) v(y) = C_\alpha \sum_y L_\alpha^{rw}(x, y) v(y) \to \mathcal{L}_\alpha v(x),$$

as $n \to \infty$ and $\varepsilon \to 0$ at a slow enough rate. This type of pointwise consistency result can be found in [37] and [12]. Furthermore, eigenvalues and eigenvectors of the graph Laplacians converge as $n \to \infty$ and $\varepsilon \to 0$ (with $\varepsilon \gg \left(\frac{\ln n}{n}\right)^{1/d}$) to eigenvalues and eigenfunctions of the corresponding Laplacian on $\mathcal{M}$, see [18]. If $\mathcal{M}$ is a manifold with boundary, then the continuum Laplacian is considered with no-flux boundary conditions. We note that the results from [18] are only stated for the case $\alpha = 0$, i.e., for the standard *random-walk Laplacian*, but the proof in [18] adapts to all $\alpha \in (-\infty, 1]$ assuming that the density ρ is smooth enough and is bounded away from zero and infinity.

Now, to understand the large sample limit of the dynamics (2.1) when $Q = Q_\alpha^{rw}$, we actually need to study the expression:

$$\sum_{x \in \mathcal{X}} u(x) Q_\alpha^{rw}(x, y), \quad y \in \mathcal{X}, \tag{4.8}$$

which in matrix form can be written as $Q_\alpha^{rw^T} u$ provided we view u as a column vector. For that purpose, we consider two smooth test functions g and h on $\mathcal{M}$. By definition of transpose,

$$\frac{1}{n} \sum_{i=1}^n h(x_i)(Q_\alpha^{rw^T} g)(x_i) = \frac{1}{n} \sum_{i=1}^n g(x_i)(Q_\alpha^{rw} h)(x_i). \tag{4.9}$$

At the continuum level, the definition of $\mathcal{L}_\alpha$ and integration by parts provide that

$$\int_{\mathcal{M}} h(x) \frac{1}{\rho(x)} \mathrm{div}_{\mathcal{M}} \left(\rho^{2(1-\alpha)} \nabla_{\mathcal{M}} \left(\frac{g}{\rho^{1-2\alpha}} \right) \right) \rho(x) dVol_{\mathcal{M}}(x)$$
$$= -\int_{\mathcal{M}} g(x) \mathcal{L}_\alpha h(x) \rho(x) dVol_{\mathcal{M}}(x)$$

By the convergence of Q_α^{rw} toward $-\mathcal{L}_\alpha$ as $n \to \infty$, we can conclude that right-hand sides converge, and thus the left-hand sides do too; notice that the $\rho(x)dx$ on both sides appear because in both sums in (4.9) the points x_i are distributed according to ρ. From this computation, we can identify the limit of (4.8) as

$$\frac{1}{\rho(y)}\text{div}_{\mathcal{M}}\left(\rho^{2(1-\alpha)}\nabla_{\mathcal{M}}\left(\frac{u}{\rho^{1-2\alpha}}\right)\right).$$

In turn, we obtain the formal continuum limit of the dynamics (2.1) when $Q = Q_\alpha^{rw}$:

$$\partial_t u_t = \frac{1}{\rho}\text{div}_{\mathcal{M}}\left(\rho^{2(1-\alpha)}\nabla_{\mathcal{M}}\left(\frac{u_t}{\rho^{1-2\alpha}}\right)\right),$$

where u_t represents the density with respect to $\rho(x)dx$. If we consider

$$f_t(x) := u_t(x)\rho(x), \tag{4.10}$$

that is, f_t is a probability density w.r.t. dx, we see it satisfies

$$\partial_t f_t = \text{div}_{\mathcal{M}}\left(\rho^{2(1-\alpha)}\nabla_{\mathcal{M}}\left(\frac{f_t}{\rho^{2(1-\alpha)}}\right)\right) = \Delta_{\mathcal{M}} f_t - 2(1-\alpha)\text{div}_{\mathcal{M}}(f_t \nabla_{\mathcal{M}} \log(\rho)), \tag{4.11}$$

where the last equality follows from an application of the product rule to the term $\nabla_{\mathcal{M}}\left(\frac{f}{\rho^{2(1-\alpha)}}\right)$. Notice that after considering a time change $t \leftarrow \frac{t}{3-2\alpha}$, we can rewrite Eq. (4.11) as

$$\partial_t f_t = (1-\beta_\alpha)\Delta_{\mathcal{M}} f_t - \beta_\alpha \text{div}_{\mathcal{M}}(f_t \nabla_{\mathcal{M}} \log(\rho)), \tag{4.12}$$

where $\beta_\alpha = (2-2\alpha)/(3-2\alpha) \in [0, 1]$.

Using the above analysis and Remark 4.1, we can also conclude that the (formal) large sample limit of equation (2.1) with $Q = Q_\beta$ and potential $B = -\frac{1}{\rho}$ is given by

$$\partial_t f_t = (1-\beta)\Delta_{\mathcal{M}} f_t - \beta \text{div}_{\mathcal{M}}(f_t \nabla_{\mathcal{M}} \log(\rho)), \tag{4.13}$$

that is, the same continuum limit as for the Fokker–Planck equations constructed using the rate matrix Q_α for α such that $\beta = \beta_\alpha$.

Remark 4.2 Notice that when $\beta = 1$, Eq. (4.13) reduces to the heat equation on $\mathcal{M}$ where no role is played by ρ. In this case, clustering is determined completely by the geometric structure of $\mathcal{M}$. On the other hand, when $\beta = 0$, Eq. (4.13) reduces to mean shift dynamics on $\mathcal{M}$ as discussed in Sect. 1.1.

Remark 4.3 Several works in the literature have established precise connections between operators such as graph Laplacians built from random data and analogous differential operators defined at the continuum level on smooth compact manifolds without boundary. For pointwise consistency results, we refer the reader to [37, 21, 20, 4, 40, 19]. For *spectral convergence* results, we refer the reader to [46] where the regime $n \to \infty$ and ε constant has been studied. Works that have studied regimes where ε is allowed to decay to zero (where one recovers differential operators and not integral operators) include [36, 5, 16, 28, 6, 13, 48]. Recent work [10] considers the spectral convergence of $\mathcal{L}_\alpha$ with self-tuned bandwidths and includes the $\alpha < 0$ range. The work [7] provides regularity estimates of graph Laplacian eigenvectors.

The case of manifolds *with* boundary has been studied in papers like [43, 39, 28, 18]. It is important to highlight that the specific computations presented in our Sect. 4.1 would have to be modified to take into account the effect of the boundary, in particular on the kernel density estimate. However, we remark that the tools and analysis from the papers mentioned above can be used to generalize these computations.

Remark 4.4 A connection between Fokker–Planck equations at the continuum level and the graph dynamics induced by Q^{rw}_α when $\alpha = 1/2$ was explicitly mentioned in [31]. To establish an explicit link between mean shift and spectral clustering, however, we need to consider the range $(-\infty, 1]$ for α. In the diffusion maps literature, the interval $[0, 1]$ is considered as natural range for α, but the analysis presented in this section explains why $(-\infty, 1]$ is in fact a more natural choice.

Remark 4.5 Besides the Fokker–Planck interpolations considered in Sect. 3.1, another family of data embeddings that are used to interpolate geometry-based and density-driven clustering algorithms is based on the path-based metrics studied in [26, 27].

4.3 *The Witten Laplacian and Some Implications for Data Clustering*

In the previous section, we presented a (formal) connection between Fokker–Planck equations on proximity graphs and Fokker–Planck equations on manifolds. In this section, we use this connection to illustrate why the Fokker–Planck interpolation is expected to produce better clusters in settings like the blue sky problem discussed in our numerical experiments in Sect. 5.2.6 where both pure mean shift and pure spectral clustering perform poorly. For simplicity, we only consider the Euclidean setting.

We start by noticing that Eq. (4.13) can be rewritten as

$$\partial_t \tilde{f}_t = -\Delta_\varrho \tilde{f}_t, \tag{4.14}$$

after considering the transformation:

$$f = \exp\left(-\frac{1-\beta}{2\beta}\varrho\right)\tilde{f}, \quad \varrho := -\log(\rho).$$

In the above, the operator Δ_ϱ is the *Witten Laplacian* (see [47] and [30]) associated with the potential $\frac{1-\beta}{2}\varrho$ which is defined as

$$\Delta_\varrho v := -\beta^2 \Delta v + \frac{(1-\beta)^2}{4}|\nabla\varrho|^2 v - \frac{\beta(1-\beta)}{2}(\Delta\varrho)v. \tag{4.15}$$

From the above, we conclude that the Fokker–Planck dynamics (4.13) can be analyzed by studying the dynamics (4.14). In turn, some special properties of the Witten Laplacian Δ_ϱ that we review next allow us to use tools from spectral theory to study equation (4.14) and in turn also the type of data embedding induced by our Fokker–Planck equations on graphs.

To begin, notice that

$$\Delta_\varrho = \left(-\beta \mathrm{div} + \frac{(1-\beta)}{2}\nabla\varrho\right)\left(\beta\nabla + \frac{(1-\beta)}{2}\nabla\varrho\right). \tag{4.16}$$

From the above, we see that $\langle \Delta_\varrho f, g\rangle_{L^2(\mathcal{M})}$ can be written as

$$\langle \Delta_\varrho g, h\rangle_{L^2(\mathcal{M})} = \int_{\mathcal{M}} \left\langle \beta\nabla g + g\frac{(1-\beta)}{2}\nabla\varrho, \beta\nabla h + h\frac{(1-\beta)}{2}\nabla\varrho\right\rangle dx,$$

from where we conclude that $\langle \Delta_\varrho f, g\rangle_{L^2(\mathcal{M})}$ is a quadratic form with associated Dirichlet energy:

$$D(f) := \int_{\mathcal{M}} \left|\nabla f + f\frac{(1-\beta)}{2}\nabla\varrho\right|^2 dx. \tag{4.17}$$

When $\mathcal{M}$ is compact, it is straightforward to show that there exists an orthonormal basis $\{\varrho_k\}_{k\in\mathbb{N}}$ for $L^2(\mathcal{M})$ consisting of eigenfunctions of Δ_ϱ with corresponding eigenvalues $0 = \lambda_1 < \lambda_2 \leqslant \lambda_2 \leqslant \dots$ that can be characterized using the Courant–Fisher minmax principle. Using the spectral theorem, we can then represent a solution to (4.14) as

$$\tilde{f}_t = \sum_{k=1}^{\infty} e^{-t\lambda_k}\langle \tilde{f}_0, \varrho_k\rangle_{L^2(\mathcal{M})}\varrho_k$$

and conclude that the dynamics (4.14) are strongly influenced by the eigenfunctions with smallest eigenvalues.

We now explain the implication of the above discussion on data clustering. Suppose that we consider a data distribution in $\mathbb{R}^2$ as the one considered in Sect. 5.2.6 modeling the blue sky problem, so that in particular it has product structure, i.e., $\rho(x, y) = \rho_1(x)\rho_2(y)$. In this case, we can use the additive structure of the potential $\varrho = -\log(\rho(x, y)) = -\log(\rho_1(x)) - \log(\rho_2(y)) =: \varrho_1(x) + \varrho_2(y)$ to conclude that the set of eigenvalues of Δ_ϱ and a corresponding orthonormal basis of eigenfunctions can be obtained from

$$\lambda_{1,i} + \lambda_{2,j}, \quad \varrho_{1,i}(x)\varrho_{2,j}(y),$$

where $(\lambda_{1,i}, \varrho_{1,i})$ are the eigenpairs for the 1D Witten Laplacian Δ_{ϱ_1} and $(\lambda_{2,j}, \varrho_{2,j})$ are the eigenpairs for Δ_{ϱ_2}. In particular, the first non-trivial eigenvalue of Δ_ϱ and its corresponding eigenfunction (which will be the effective discriminators of the two desired clusters if λ_3 is considerably larger than λ_2) are either $\lambda_{1,2}$ and $\varrho_{1,2}(x)\varrho_{2,1}(y)$ or $\lambda_{2,2}$ and $\varrho_{1,1}(x)\varrho_{2,2}(y)$. This discussion captures the competition between a horizontal and a vertical partitioning of the data in the context of the blue sky problem from Sect. 5.2.6. While we are not able to retrieve the desired horizontal partitioning by setting $\beta = 0$ or $\beta = 1$, we can identify the correct clusters by setting β strictly between zero and one (closer to one than to zero). We notice that the results from [30] can be used to obtain precise quantitative information on the small eigenvalues of the 1D Witten Laplacians Δ_{ϱ_1} and Δ_{ϱ_2} when β is close to one (i.e., the diffusion term is small), which we can use to determine whether $\lambda_{2,2} < \lambda_{1,2}$ or vice versa.

5 Numerical Examples

We now turn to the details of our numerical method and examples illustrating its properties. We begin, in Sect. 5.1, by describing the details of our numerical approach. We provide Algorithm 5.1 for its practical implementation.

In Sect. 5.2, we consider several numerical examples, beginning with examples in one spatial dimension. In Fig. 4, we illustrate how the graph dynamics for the transition rate matrices Q_β and Q_α^{rw} can be visualized as the evolution of a continuum density, and in Fig. 5, we illustrate the good agreement between the graph dynamics and the dynamics of the corresponding continuum Fokker–Planck equation. In Fig. 6, we show how the clustering performance of our method depends on the balance between drift and diffusion (β), the time of clustering (t), and the number of clusters (k); we also illustrate the benefits and limitations of using the energy of the k-means clustering to identify the number of clusters. In Fig. 7, we consider the role of the kernel density estimate in clustering dynamics, showing how adding diffusion to mean shift dynamics can help the dynamics overcome spurious local minimizers in the kernel density estimate, leading to better clustering performance. In Fig. 8, we illustrate the interplay between the underlying data distribution and the balance between drift and diffusion (β) .

Next, we consider several examples in two dimensions. In Figs. 10 and 11, we consider a model of the *blue sky problem*, in which data points are distributed over two elongated clusters that are separated by a narrow low-density region. We illustrate how diffusion dominant dynamics prefer to cluster based on the geometry of the data, leading to poor performance. Similarly, pure mean shift dynamics can exhibit poor clustering due to local maxima in the kernel density estimate. By interpolating between the two extremes, we observe robust clustering performance, for a wide range of graph connectivity (ε). Finally, in Figs. 12 and 13, we consider an example in which three blobs are connected by two bridges: one wide, low-density bridge and another narrow, high-density bridge. This example is constructed so that there is no correct clustering into two clusters. Instead, a geometry-based clustering method would prefer to cut the thin bridge, and a density-based clustering method would prefer to cut the wide bridge. We show how varying the balance between drift and diffusion in our method (β) allows our method to cut either bridge.

5.1 Numerical Method

For our numerical experiments, we consider a domain $\Omega \subseteq \mathbb{R}^d$ and a density $\rho : \Omega \to [0, +\infty)$ normalized so that $\int_\Omega \rho = 1$. All PDEs on Ω will be considered with no-flux boundary conditions, as the solutions of the graph-based equations converge to the solutions of PDE with no-flux boundary conditions (observed in [12] and rigorously proved in [18] for Laplacians).

We draw n samples $\{x_i\}_{i=1}^n$ from ρ on Ω. These samples are the nodes of our weighted graph, and for all simulations, the weights on the graph are given by a Gaussian weight function

$$w(x_i, x_j) = \varphi_\varepsilon(|x_i - y_j|), \quad \varphi_\varepsilon(a) = \frac{e^{-a^2/2\varepsilon^2}}{(2\pi\varepsilon^2)^{d/2}}, \quad a \in \mathbb{R}. \tag{5.1}$$

In our one-dimensional simulations, we take the graph bandwidth parameter ε to be

$$\varepsilon = \sqrt{2} \max_i \min_{j:j \neq i} |x_i - x_j|; \tag{5.2}$$

that is, ε equals the maximum distance to the closest node. We note that even in higher dimensions, the ε above scales as $(\ln n/n)^{1/d}$ with the number of nodes n. This has been identified as the threshold, in terms of n, at which the graph Laplacian is spectrally consistent with the manifold Laplacian [18]. In Fig. 11, we illustrate how the choice of ε impacts dynamics and, ultimately, clustering performance.

With this graphical structure, we now recall the weighted diffusion transition rate matrix Q_α^{rw}, for $\alpha \in (-\infty, 1]$, as in Eq. (1.12), with the constant $C_\alpha = ((3 - 2\alpha)\varepsilon^2)^{-1}$,

$$Q_\alpha^{rw}(x, y) := \frac{1}{(3 - 2\alpha)\varepsilon^2} \begin{cases} \frac{w_\alpha(x,y)}{\sum_{z \neq x} w_\alpha(x,z)} & \text{if } x \neq y, \\ -1 & \text{if } x = y, \end{cases} \tag{5.3}$$

$$w_\alpha(x, y) := \frac{w(x, y)}{d(x)^\alpha d(y)^\alpha}, \qquad d(x_i) = \sum_{x_j \neq x_i} w(x_i, x_j). \tag{5.4}$$

Similarly, we recall the transition rate matrix Q_β, for $\beta \in [0, 1]$, as in Eq. (3.1), with the constant $C_{ms} = (\varepsilon^2 n)^{-1}$,

$$Q_\beta := \beta Q^{ms} + (1-\beta) Q_1^{rw}, \tag{5.5}$$

$$Q^{ms}(x,y) := \frac{1}{\varepsilon^2 n} \begin{cases} \left(-\frac{1}{\hat{\rho}_\delta(y)} + \frac{1}{\hat{\rho}_\delta(x)}\right)_+ w(x,y), & \text{for } x \neq y, \\ -\sum_{z \neq x} \left(-\frac{1}{\hat{\rho}_\delta(z)} + \frac{1}{\hat{\rho}_\delta(x)}\right)_+ w(x,z), & \text{for } x = y, \end{cases} \tag{5.6}$$

$$\hat{\rho}_\delta(x) := \frac{1}{n} \sum_{y \in \mathcal{X}} \varphi_\delta(x-y). \tag{5.7}$$

Unless otherwise specified, we take the bandwidth δ in our kernel density estimate for our one-dimensional examples to be

$$\delta = \sqrt{2} \left(\frac{|\Omega|}{n}\right)^{0.5}. \tag{5.8}$$

With these transition rate matrices in hand, we may now consider solutions u_t of (2.1) when $Q = Q_\alpha^{rw}$ or when $Q = Q_\beta$. We solve the ordinary differential equations describing the graph dynamics by directly computing the matrix exponential e^{tQ} in each case; see Definition 2.1. Following the discussion in Sect. 4.2, we know that for each of these dynamics, as $n \to +\infty$ and $\varepsilon, \delta \to 0$ (at an n dependent rate that is not too fast), the measures $\sum_{j=1}^n u_t(x_j)\delta_{x_j}$ are expected to converge to solutions f_t of the following Fokker–Planck equation:

$$\partial_t f_t = (1-\beta)\Delta f_t - \beta \operatorname{div}(f_t \nabla \log(\rho)), \tag{5.9}$$

where for the Q_α^{rw} dynamics, we take

$$\beta = \beta_\alpha = (2-2\alpha)/(3-2\alpha) \tag{5.10}$$

The steady state of the equation is the corresponding Maxwellian distribution

$$c_{\rho,\beta}\, \rho^{\beta/(1-\beta)}(x), \tag{5.11}$$

where $c_{\rho,\beta} > 0$ is a normalizing constant chosen so that the distribution integrates to one over Ω. Note that, if $d(x_i)$ represents the degrees of the graph vertices, as in Eq. (1.8), then the function $u_t(x_i)d(x_i)$ likewise converges to $f_t(x)$ as the number of nodes in our sample $n \to +\infty$. Consequently, when comparing our graph dynamics to the PDE dynamics, we will often plot $u_t(x_i)d(x_i)$ and $f_t(x)$.

Finally, we use the embedding maps $\hat{\Psi}_\alpha$ and $\hat{\Psi}^\beta$ from Sects. 2.1 and 3.1 to cluster the nodes. In particular, we apply k-means to the vectors $\{\hat{\Psi}_\alpha(x_i)\}_{i=1}^n$ and $\{\hat{\Psi}^\beta(x_i)\}_{i=1}^n$, obtaining in this way a series of maps from nodes $\{x_i\}_{i=1}^n$ to cluster centers $\{l_m\}_{m=1}^{k}$. Nodes mapped to the same cluster center are identified as belonging to the same cluster. While we will not discuss at any depth the methods to select the best number of clusters, we note that a number of methods to do so (in particular the elbow method and the gap statistics [38]) rely on the value of the k-means energy,

$$E_k = \frac{1}{n}\sum_{i=1}^{n} \min_{m=1,\dots k} |\Psi(x_i) - l_j|^2, \tag{5.12}$$

for each relevant Ψ. Note that E_k always decreases with k. While a large decrease in the energy as k increases is indicative of the improved approximation of data by cluster centers, the size of the jumps is truly telling only if we compare it with the relevant model for the data considered, see [38] and discussion in Sect. 5.2.3. For ease of visualization, in our numerical examples, we will plot the *normalized* k-means energy, which is rescaled so that energy of a single cluster equals one,

$$E_k^{\text{norm}} = E_k/E_1. \tag{5.13}$$

All of our simulations are conducted in Python, using the Numpy, SciPy, Sci kit-learn, and MatPlotLib libraries [42, 24, 23, 33]. In particular, we use the Sci kit-learn implementation of k-means to cluster the embedding maps.

Algorithm 1 Dynamic clustering algorithm for Q_β or Q_α^{rw}

```
**Input:** $\{x_i\}_{i=1}^{n}$, $\varepsilon$, $\delta$, $t$, $k$
$Q = Q_\beta$ or $Q = Q_\alpha^{rw}$
$\hat{\Psi}_Q(x_i) = (e^{tQ})_{(i,j=1,\dots n)}$ for $i = 1, \dots, n$
$l_m$ = Kmeans.fit($\hat{\Psi}_Q(x_1), \dots, \hat{\Psi}_Q(x_n)$) with $\text{n}_{\text{clusters}} = k$
```

5.2 *Simulations*

We now turn to simulations of the graph dynamics, PDE dynamics, and clustering.

5.2.1 Graph Dynamics as Density Dynamics

In Fig. 4, we illustrate how the dynamics on a graph can be visualized as the evolution of a density on the underlying domain $\Omega = [-1.5, 1.5]$. The right column of Fig. 4 illustrates two choices of data density (blue line),

$$\rho_{\text{two bump}}(x) = 4c\varphi_{0.5}(x + 0.5) + c\varphi_{0.25}(x - 1.25)$$

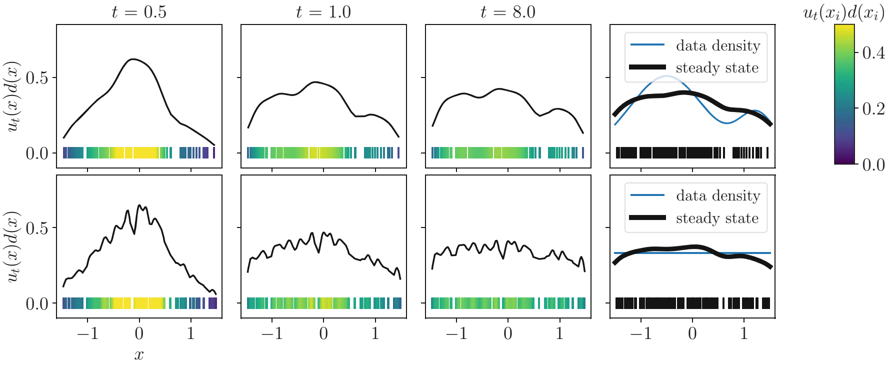


Fig. 4 Illustration of the graph dynamics u_t for Q_β, $\beta = 0.25$, from initial condition δ_{x_i}, $x_i = -0.1$, for two choices of data density: $\rho_{\text{two bump}}$ (top) and ρ_{uniform} (bottom). The first three columns show the evolution of $u_t(x)d(x)$ at times $t = 0.1, 0.5$, and 8.0, with the color of the markers representing the value of $u_t(x_i)d(x_i)$ at each node. The last column depicts the data density (blue line) from which the nodes of the graph $\{x_i\}_{i=1}^n$ (black markers) are sampled, as well as the steady state of the dynamics (thick black line)

$$\text{and} \qquad \rho_{\text{uniform}}(x) = \frac{1}{3}, \qquad x \in \mathbb{R}. \tag{5.14}$$

The constant $c > 0$ is chosen so that the integral of both densities over the domain equals one. We sample the nodes of the graph $\{x_i\}_{i=1}^n$ (black markers) from each density, with $n = 147$ nodes sampled for $\rho_{\text{two bump}}$ and $n = 140$ nodes sampled for ρ_{uniform}. The first three columns show the evolution of the graph dynamics $u_t(x)d(x)$ from Eq. (2.1) for $Q = Q_\beta$ with $\beta = 0.25$ and initial condition δ_{x_i}, $x_i = -0.1$, where the top row corresponds to the graph arising from $\rho_{\text{two bump}}$ and the bottom row corresponds to the graph arising from ρ_{uniform}. The color of the markers represents the value of $u_t(x_i)d(x_i)$ at each node. We observe in both rows that $u_t(x)d(x)$ approaches the steady state of the corresponding continuum PDE (5.11), depicted in a thick black line in the fourth column.

The fact that $d(x)u_t(x)$ appears more jagged in the bottom row compared to the top row is due to the smaller value of ε in the graph weight matrix: see Eqs. (5.1–5.2). Since our sample of the data density in the top row has an isolated node at $x_i = 1.44$, this leads to a significantly larger value of ε in the simulations on the top row ($\varepsilon = 0.13$), compared to the bottom row ($\varepsilon = 0.03$).

5.2.2 Comparison of Graph Dynamics and PDE Dynamics

In Fig. 5, we compare the graph dynamics to the corresponding Fokker–Planck equation (5.9). We consider the data density given by $\rho_{\text{two bump}}$ and initial condition δ_{x_i}, for $x_i = -0.1$. The graphs are built from $n = 625$ samples of the data density, and solutions are plotted at times $t = 0.5, 1.0, 8.0$.

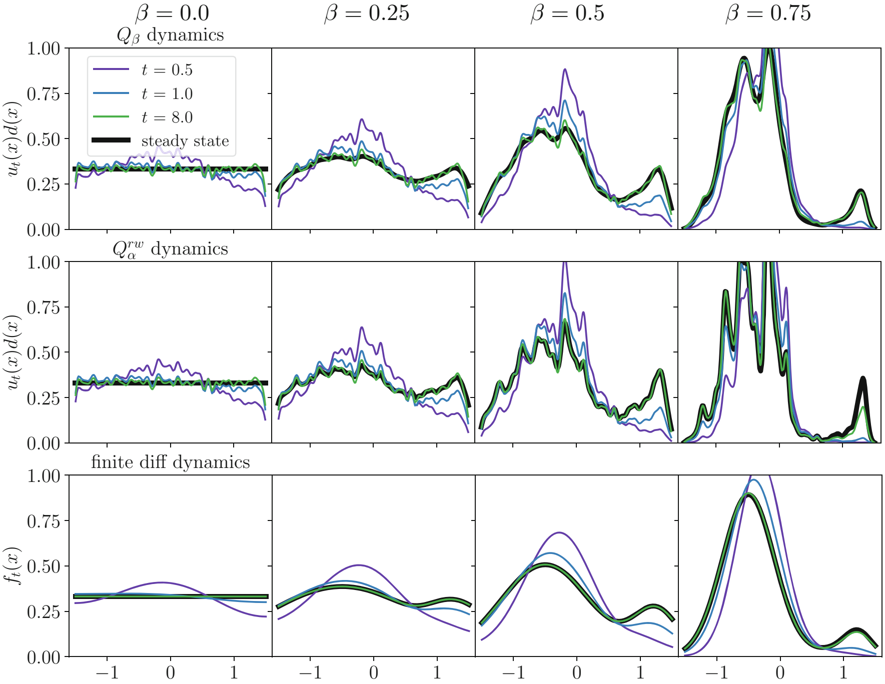


Fig. 5 Comparison of the graph dynamics for Q_β (top) and Q_α (middle) with the PDE dynamics (bottom). The data density is $\rho_{\text{two bump}}$, and the initial data is δ_{x_i} for $x_i = -0.1$. The graphs are built from $n = 625$ samples of the data density. The steady states are obtained from equation (5.15) for the graph dynamics and equation (5.11) for the finite difference dynamics

The top row illustrates the graph dynamics $u_t(x)d(x)$ arising from the transition rate matrix Q_β, for $\beta = 0, 0.25, 0.5, 0.75$. The middle row illustrates $u_t(x)d(x)$ arising from Q_α^{rw} for $\alpha = 1.0, 0.83, 0.5, -0.5$. (The values of α are chosen to give the same balance between drift and diffusion as in the top row; see equation (5.10).) The last row shows a finite difference approximation of the Fokker–Planck equation (5.9). We compute solutions of the PDEs using a semi-discrete, upwinding finite difference scheme on a one-dimensional grid, with 200 spatial gridpoints and continuous time. This reduces the PDEs to a system of ODEs, which we then solve using the SciPy odeint method.

The steady states we plot for the graph dynamics are given by the following equation:

$$c_{n,\delta,\beta}(\hat{\rho}_\gamma(x))^{\beta/(1-\beta)}, \quad \hat{\rho}_\gamma(x) = \frac{1}{n}\sum_{y\in\mathcal{X}} \psi_\gamma(x-y), \quad \psi_\gamma(x) = \frac{1}{(2\pi)^{1/2}\gamma} e^{-|x|^2/(2\gamma^2)}, \tag{5.15}$$

where $c_{n,\delta,\beta}$ is a normalizing constant chosen, so the steady state integrates to one over Ω. For the Q_β dynamics, we choose the standard deviation $\gamma = \delta$, and for the Q_α^{rw} dynamics, we choose $\gamma = \varepsilon$. Recall that $\hat{\rho}_\delta$ is the kernel density estimator used in the construction of the transition matrix Q^β; see Eq. (2.7). The steady states for the PDE dynamics are given by Eq. (5.11).

Interestingly, even though there is no explicit kernel density estimate of the data in the construction of the transition rate matrix Q_α^{rw}, the above simulations demonstrate better agreement of these dynamics as $t \to +\infty$ with the steady state arising from a kernel density estimate (5.15) than with the steady state arising directly from the data density (5.11). This can be seen by observing the good agreement at time $t = 8.0$ with the solid black line shown in the middle row, rather than the solid black line shown in the bottom row. This suggests that the Q_α^{rw} operator effectively takes a KDE of the data density with bandwidth $\varepsilon > 0$, corresponding to the scaling of the weight matrix on the graph.

5.2.3 Clustering Dynamics

In Fig. 6, we illustrate how the graph dynamics u_t of the transition rate matrix Q_β can be used for clustering. The underlying data density is $\rho_{\text{two bump}}$, from which we choose $n = 204$ samples. We consider $\beta = 0.25, 0.9$, and 1.0, corresponding to the three columns of the figure. The top portion of the figure shows the results of the k-means clustering algorithm for $k = 2, 3, 4$. Each plot depicts the data samples at times $t = 10^{-1}, 1, 10$, coloring the samples according to which cluster they belong. The top right panel on the figure shows the data distribution and the kernel density estimate of the data distribution, which is used to construct the transition rate matrix Q_β. The bottom of the figure shows the value of the k-means energy E_k (5.13) for each clustering normalized so that $k = 1$ clustering (all nodes in a single cluster) has energy $E_1 = 1$.

For all β and k, there is poor clustering behavior early in time, $t = 0.1$, suggesting that the Fokker–Planck dynamics have not had time to effectively mix within clusters. This can be seen by comparing the colors of the nodes to the data distribution displayed on the right: a correct clustering should identify one cluster for the large bump and another cluster for the small bump. This can also be seen by considering the k-means energy, which is largest at $t = 0.1$, and shows little variation for different choices of k.

On the other hand, we observe the best clustering performance for $\beta = 0.9$ and time $t = 10$. Examining the colors of the nodes for $k = 2$ reveals that the correct clusters are found. Furthermore, this clustering remains fairly stable as k is increased. This can also be seen in the k-means energy, which shows a substantial decrease from $k = 1$ to $k = 2$, but remains stable for $k = 3, 4$, suggesting that two clusters are the correct number of clusters.

While $\beta = 0.25$ and $\beta = 1$ do not offer good clustering performance, they do shed light on key properties of our method, once time is sufficiently large to have allowed the dynamics to effectively mix, $t = 10$. For example, when $\beta =$

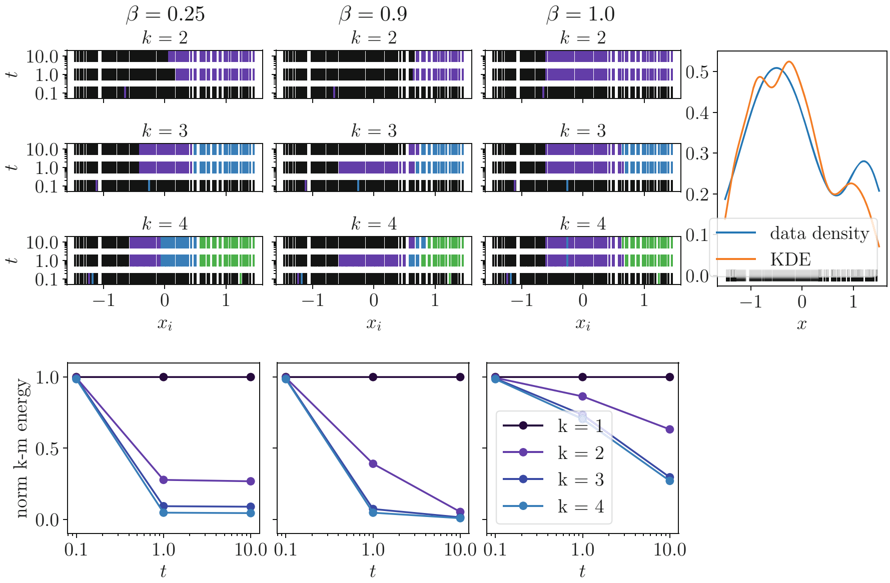


Fig. 6 Clustering performance of the graph dynamics for the transition rate matrix Q_β. The top portion of the figure shows the results of the k-means clustering algorithm for $k = 2, 3, 4$ (rows) and $\beta = 0.25, 0.9, 1.0$ (columns). The color of a node indicates the cluster to which it belongs. On the top right, we show the data density from which $n = 204$ nodes are sampled and the KDE of the data density used to construct Q_β. The bottom of the figure shows the value of the normalized k-means energy

0.25, diffusion dominates the dynamics, so that the density of the data distribution does not play a strong role in clustering. In fact, we see that the clusters are almost entirely driven by the geometry of the data distribution, which is fairly uniform on the domain: when $k = 2$, the clusters are essentially even halves of the domain; when $k = 3$, they are even thirds; and when $k = 4$, they are even quarters. The lack of awareness of density when $\beta = 0.25$ inhibits correct cluster identification.

We observe the opposite problem when $\beta = 1$. In this case, the dynamics are driven entirely by density, with no diffusion. However, the density driving the dynamics is not the exact data density, but the kernel density estimate. Due to noise in the KDE, an artificial local minimum appears near $x = -0.75$, causing $k = 2$ to cluster the nodes to the left and right of this local minimum and causing $k = 3$ to cluster the nodes into three even groups, separated by the two local minima of the KDE. Unlike in the case $\beta = 0.9$, when $\beta = 1.0$, there is no diffusion to help the dynamics overcome spurious local minima in the KDE, leading to inferior clustering performance.

We close by considering the role of the k-means energies in identifying the correct number of clusters. First, consider the case of a uniform data distribution. In this case, the k-means energy for $k = 2$ would be $\frac{1}{4}$ and for $k = 3$ would be $\frac{1}{9}$.

Consequently, while the correct number of clusters for the uniform data distribution is one, the k-means energy still drops significantly as k is increased. For this reason, we caution that looking for the largest drop of the energy alone is not a good criterion for determining the correct number of clusters. Determining the correct number of clusters remains an active area of research, including, for example, the study of gap statistics [38], in which the energy is compared to the energy one would have if the data were uniform.

5.2.4 Effect of the Kernel Density Estimate on Clustering

Figure 7 illustrates the effect that the bandwidth δ of the kernel density estimate of the data distribution has on clustering, see Eq. (2.7). The data distribution is given by a piecewise constant function, shown in the rightmost column. The number of samples chosen is $n = 680$, and the clustering is performed at time $t = 30$. The graph connectivity parameter ε is chosen as in Eq. (5.2), equaling 0.015. The top two rows show clustering performance for Q_β for $\beta = 0.25, 0.9, 1$, and the bottom

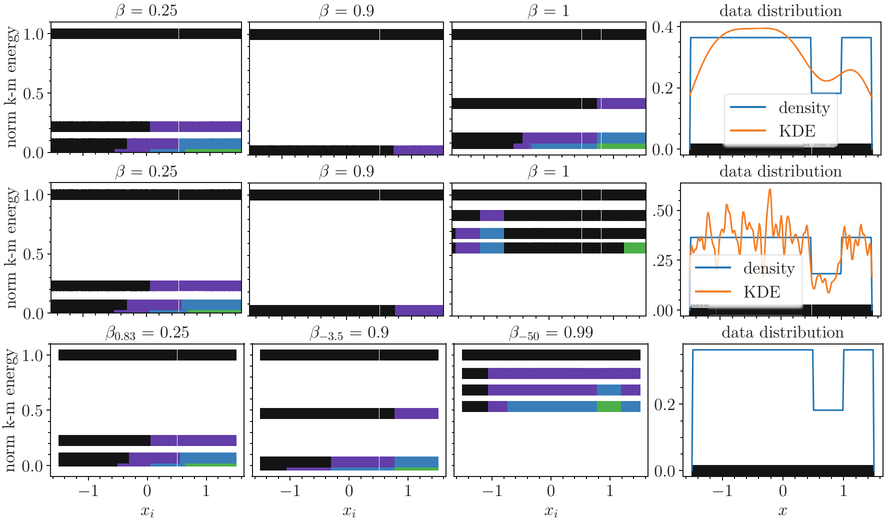


Fig. 7 Effect of the bandwidth of the kernel density estimator on clustering, for $n = 676$ samples clustered at time $t = 30$. First two rows show the clustering with Q_β, (5.5) for KDE bandwidths $\delta = 0.2$ and $\delta = 0.015$, while the third shows the dynamics of Q^{rw}_α, (5.3). The first three columns show clustering performance for different balances of drift and diffusion, and the fourth column shows the data distribution and kernel density estimate. Note that, since no explicit kernel density estimate is used in the construction of Q^{rw}_α, none is shown in the third row. The colors of the samples indicate the clusters to which they belong, and the height of the samples in each frame indicates the value of the normalized k-means energy (5.12). The top row of markers in each frame corresponds to a single cluster ($k = 1$), the next one represents two clusters ($k = 2$), then three and four clusters

row shows clustering performance for Q_α^{rw} for $\alpha = 0.83, -3.5, -50$, where the values of α are chosen to give a comparable balance between drift and diffusion at the level of the continuum PDE; see Eqs. (5.9–5.10).

The first three columns show the clustering results for $\beta = 0.25, 0.90, 1.00$. The color of a marker indicates the cluster to which it belongs, and the height of the marker in the frame represents the value of the normalized k-means energy (5.13). Since the normalized k-means energy is decreasing in k, the top row of markers in each frame corresponds to a single cluster ($k = 1$), the next one represents two clusters ($k = 2$), then three and four clusters.

In the top row, the bandwidth of the kernel density estimate used to construct Q_β is $\delta = 0.20$, and in the middle row, $\delta = 0.015$. The effect of the bandwidth on the kernel density estimate can be seen in the rightmost column: the larger value of δ in the top row leads to a more accurate estimator of the data density than the smaller value of δ in the middle row. As no explicit kernel density estimate is used to construct the transition rate matrix Q_α^{rw}, no estimator is shown in the rightmost column of the bottom row. However, our previous numerical simulations in Fig. 5 suggest that the dynamics of Q_α^{rw} most closely match the continuum Fokker–Planck equation with a steady state induced by a kernel density estimate with bandwidth $\delta = \varepsilon$ (5.15). This is the motivation behind our choice of $\delta = 0.015 = \varepsilon$ in the second row, since it provides the closest comparison between the clustering dynamics of Q^β and Q_α^{rw}. Finally, we note that the choice of δ we suggest in Eq. (5.8) would lead to the choice $\delta = 0.07$, which is between the values considered in the top and middle rows of the figure, and leads to very similar performance for $\beta < 1$.

In the top row, when the bandwidth in the KDE is large, we observe good clustering performance for $\beta = 0.9$ and 1.0 and $k = 2$. On the other hand, $\beta = 0.25$ performs poorly, since the large amount of diffusion causes the dynamics to ignore the changes in relative density and cluster based on the fairly uniform geometry of the sampling. In the middle row, when the bandwidth in the KDE is small, we still observe good performance for $\beta = 0.9$, though $\beta = 1.0$ clusters poorly: without diffusion, the dynamics cluster based on spurious local maxima. As before, $\beta = 0.25$ identifies incorrect clusters, since it lacks information about density. Finally, as we expected, the clustering performance in the bottom row is similar to the middle row, due to the fact that the bandwidth in the middle row was chosen to match the bandwidth of the implicit kernel density estimate which appears to drive the dynamics of Q_α^{rw}. Note that, for the bottom row, the only way to increase the bandwidth of the implicit kernel density estimate would be to increase the graph connectivity parameter ε, which, for compactly supported graph weights, would lead to a more densely connected graph and thus higher computational cost.

5.2.5 Effect of Data Distribution on Clustering

Figure 8 illustrates the effect that different choices in data distribution have on the clustering method based on Q_β, for $n = 160$ nodes and at time $t = 30$. Each row considers a different data distribution: ρ_{twobump} (see Eq. (5.14)) and

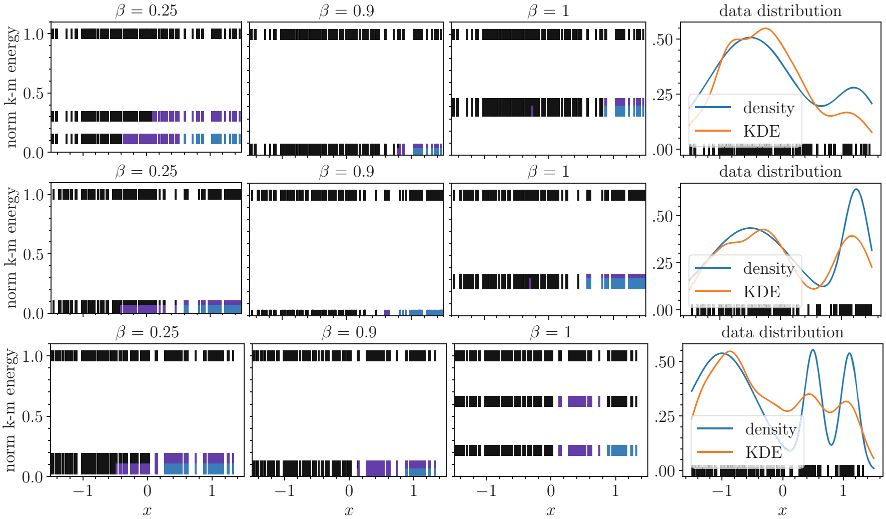


Fig. 8 Illustration of the effect that different choices of data distribution have on the clustering method based on Q_β for $n = 160$ samples at time $t = 30$. The first three columns illustrate different choices of β, where the color of a marker indicates the cluster it belongs to for $k = 1, 2, 3$, and the height of the marker in the frame represents the value of the normalized k-means energy

$$\rho_{\text{deep valley}}(x) = 7c_0\varphi_{0.5}(x + 0.5) + 3c_0\varphi_{0.15}(x - 1.25),$$

$$\rho_{\text{three bump}}(x) = c_1\varphi_{0.1}(x - 0.5) + c_1\varphi_{0.1}(x - 1.1) + 4c_1\varphi_{0.4}(x + 1),$$

where $c_0, c_1 > 0$ are normalizing constants chosen, so the densities integrate to one on $\Omega = [-1.5, 1.5]$. The right column shows the data density, the sample of $n = 160$ nodes, and the kernel density estimate used to construct the transition rate matrix Q_β. The first three columns show the clustering results for $\beta = 0.25, 0.90, 1.00$. The color of a marker indicates the cluster to which it belongs for $k = 1, 2, 3$, and the height of the marker in the frame represents the value of the normalized k-means energy (5.13).

In the top row, for $\rho_{\text{two bump}}$, we observe good clustering performance for all $\beta \geqslant 0.9$ and poor performance for $\beta = 0.25$: due to the good behavior of the KDE for this data distribution, problems do not arise as $\beta \to 1$, and as usual, $\beta = 0.25$ suffers due to the dominance of diffusion. In the middle row, for $\rho_{\text{deep valley}}$, we again observe good performance for all $\beta \geqslant 0.9$, and we even observe good performance for $\beta = 0.25$ when $k = 2$. This is due to the sparse sampling at the deep valley, which leads to a change in the geometry of the nodes: a gap that even diffusion dominant dynamics can detect. Finally, in the bottom row, for $\rho_{\text{three mountains}}$, we observe good performance for $\beta \geqslant 0.9$. Again, $\beta = 0.25$ is

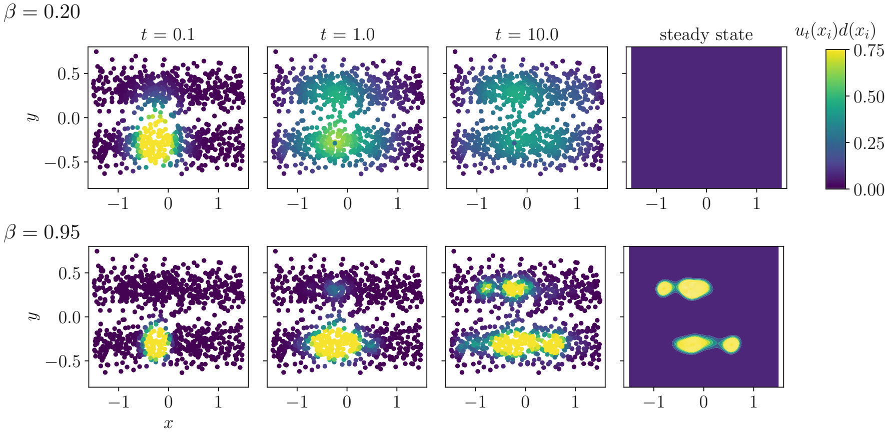


Fig. 9 Graph dynamics of Q_β on $\rho_{\text{blue sky}}$ for $n = 965$ samples for $\beta = 0.20$ (top) and $\beta = 0.95$ (bottom). The initial condition is δ_{x_i,y_i} for $(x_i, y_i) = (-0.26, -0.29)$. In the first three columns, the dots represent the locations of the samples, and the colors of the markers represent the value of $u_t(x_i)d(x_i)$. In the fourth column, we plot the steady state of the corresponding continuum PDE (5.15)

able to capture some correct information when $k = 2$, due to the sparsity of the data near the left valley, but it fails at the most relevant $k = 3$.

5.2.6 Blue Sky Problem

In Fig. 9, we consider the graph dynamics of Q_β on a two-dimensional data distribution inspired by the blue sky problem from image analysis, for $n = 965$ samples on the domain $\Omega = [-1.5, 1.5] \times [-1, 1]$. We choose $\varepsilon = 0.04$ and $\delta = 0.10$, in order to optimize agreement between the discrete dynamics and the steady state of the continuum PDE (5.15).

In simple terms, the blue sky problem can be described as a setting in which data points are distributed over two elongated clusters that are separated by a narrow low-density region. For concreteness, we model this with a density of the form:

$$\rho_{\text{blue sky}}(x, y) = \varphi_{1.0}(x) * (\varphi_{0.09}(y - 0.32) + \varphi_{0.09}(y + 0.32)).$$

In the top row, we choose $\beta = 0.20$, and in the bottom row, we choose $\beta = 0.95$. In both cases, we choose the initial condition for the dynamics to be δ_{x_i,y_i} for $(x_i, y_i) = (-0.26, -0.29)$. As in Fig. 4, the markers in the first four columns represent the locations of the samples, which form the nodes of our graph, and the colors of the markers represent the value of $u_t(x_i)d(x_i)$ at each node. In the rightmost column, we plot the steady state of the continuum PDE (5.15). We observe good agreement between the graph dynamics and the steady state by time $t = 10.0$. In the case

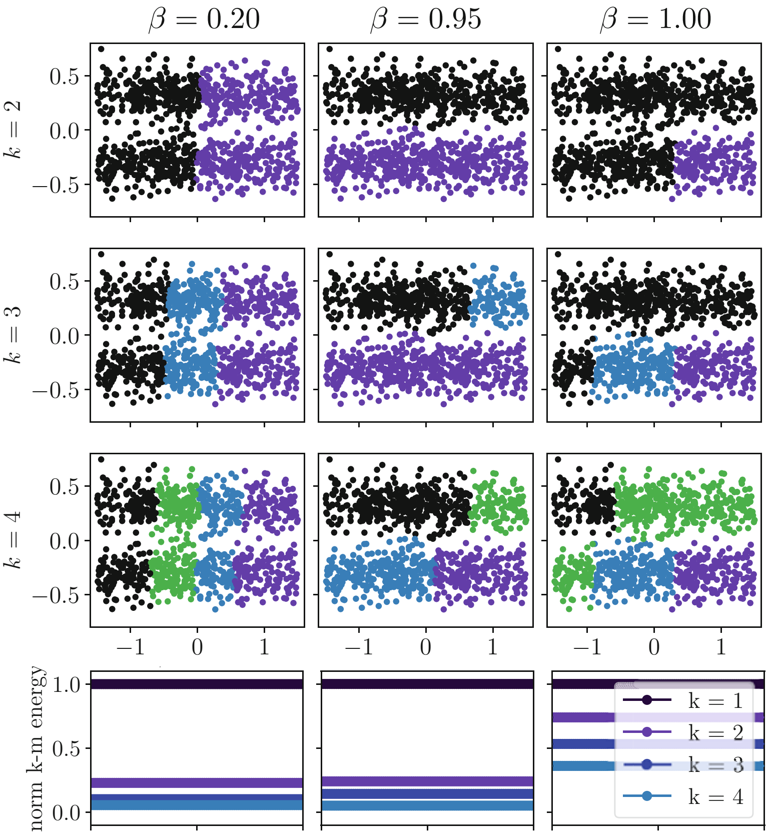


Fig. 10 Illustration of the clustering behavior of Q_β on $\rho_{\text{blue sky}}$ for $n = 965$ samples at time $t = 10$. The first three rows show the clustering behavior for $k = 2, 3, 4$, with each node colored according to which cluster it belongs. The fourth row shows the normalized k-means energy for each number of clusters k

$\beta = 0.95$, the diffuse profile of the steady state illustrates that there is significant diffusion, in spite of the fact that β is close to one.

In Fig. 10, we show the clustering behavior of Q_β on $\rho_{\text{blue sky}}$ for $n = 965$ samples at time $t = 10$. The columns correspond to $\beta = 0.2, 0.95, 1.0$. The first three rows show the clustering behavior for $k = 2, 3, 4$, with each node colored according to which cluster it belongs. In the fourth row, we show the normalized k-means energy for each choice of β.

We observe the best clustering performance for $\beta = 0.95$ and $k = 2$. Furthermore, in this case, the plot of the normalized k-means energy indicates that higher values of k do not lead to significant decreases of the energy, providing further evidence that $k = 2$ is the correct number of clusters. The clustering performance deteriorates for both $\beta = 0.2$ and $\beta = 1.0$. In the case of $\beta = 0.2$, diffusion dominates and the clustering is based on the geometry of the sample

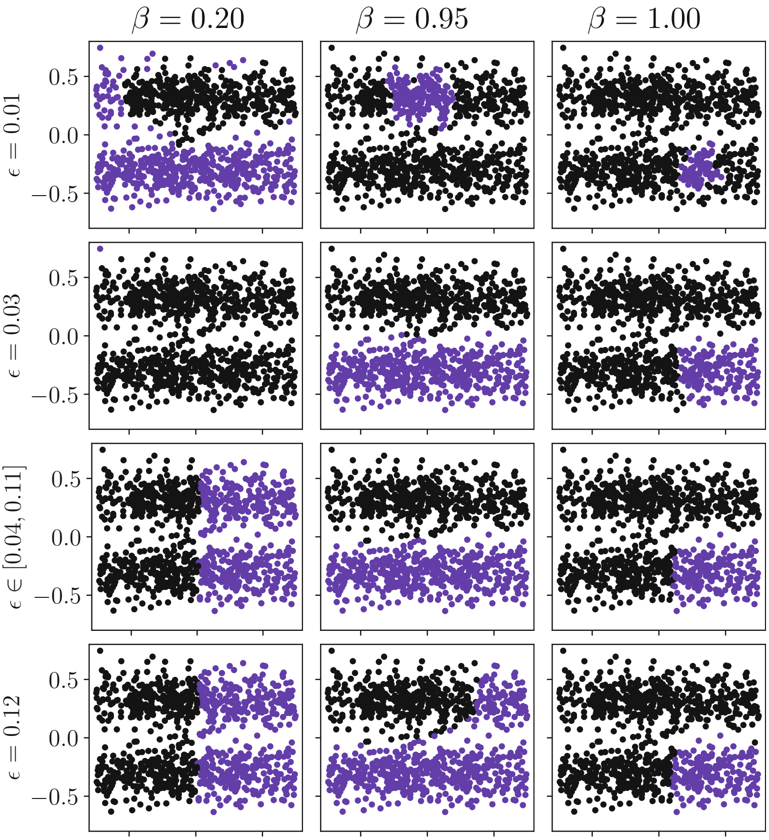


Fig. 11 For the same data distribution as in Fig. 10, we investigate how the clustering behavior of Q_β depends on the graph connectivity length scale ε for $\beta = 0.20, 0.95, 1.00$

points, preferring to cluster by slicing the sample points evenly in two pieces via the shortest cut through the data set. In the case of $\beta = 1.0$, we expect that inaccuracies in the kernel density estimation lead to spurious local minima, and in the absence of diffusion to help overcome these local minima, incorrect clusters are found. Note that simply increasing the bandwidth δ of kernel density estimate in this case would not necessarily improve performance for $\beta = 1.0$, since for a large enough bandwidth, the two lines would merge into one line.

Finally, in Fig. 11, we investigate how the clustering behavior of Q_β for $\rho_{\text{blue sky}}$ depends on the graph connectivity ε; see Eq. (5.1). The columns correspond to $\beta = 0.20, 0.95, 1.00$ and the rows correspond to $\varepsilon = 0.01, 0.03, [0.04, 0.11]$, and 0.12. We note that for a wide range of ε, the diffusion-dominated regime $\beta = 0.2$ prefers to make shorter cuts even over parts of the domain where data are dense, which is undesirable for the data considered. On the other hand, the pure mean shift suffers, as in other examples, from the tendency to

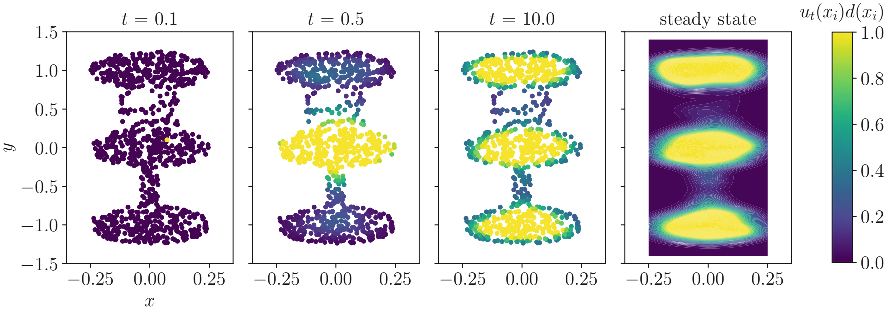


Fig. 12 Illustration of graph dynamics of Q_β on $\rho_{\text{three blobs}}$ for $n = 966$ samples, with $\beta = 0.7$ and initial condition δ_{x_i,y_i} for $(x_i, y_i) = (0.07, 0.10)$. The markers in the first three columns represent the locations of the samples, and the colors of the markers represent the value of $u_t(x_i)d(x_i)$. In the right column, we plot the steady state of the corresponding continuum PDE (5.15)

identify spurious local maxima of KDE as clusters. We observe the best clustering performance over the wide range of ε for $\beta = 0.95$. Considered together with other experiments, this suggests that adding even a small amount of diffusion goes a long way toward correct clustering.

5.2.7 Density vs. Geometry

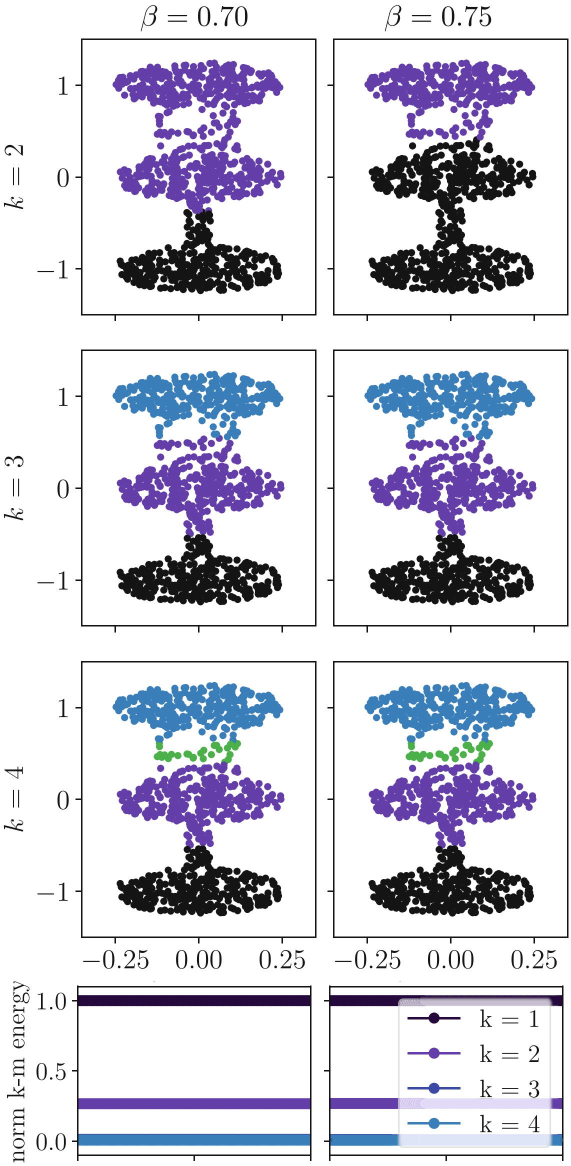
In Fig. 12, we consider the graph dynamics of Q_β on a two-dimensional data distribution chosen to illustrate how the competing effects of density and geometry depend on the parameter β. We choose $n = 966$ samples, $\varepsilon = 0.07$, and $\delta = 0.05$, in order to optimize agreement between the discrete dynamics and the continuum steady state.

The data density, which we refer to as $\rho_{\text{three blobs}}$, is given by a piecewise constant function that is equal to height one on the three circles of radius 0.25, as well as on the wide rectangle $[0.25, 0.75] \times [-0.125, 0.125]$ on the top. On the narrow rectangle $[-0.75, 0.25] \times [-0.04, 0.04]$ on the bottom, the piecewise constant function has height four. Finally, the data density is multiplied by a normalizing constant so that it integrates to one over the domain $\Omega = [-1.5, 1.5] \times [-1, 1]$.

In Fig. 12, we choose $\beta = 0.7$ and initial condition for the dynamics to be δ_{x_i,y_i} for $(x_i, y_i) = (0.07, 0.10)$. The locations of the markers represent the samples from the data distribution, and the colors of the markers represent the value of $u_t(x_i)d(x_i)$ at each node. In the right column, we plot the steady state of the corresponding continuum PDE (5.15). We observe good agreement with the graph dynamics and the steady state by time $t = 10$.

In Fig. 13, we show the clustering behavior of the Q_β on $\rho_{\text{three blobs}}$ for $n = 966$ samples at time $t = 10$. The two columns correspond to $\beta = 0.7$ and 0.75. The

Fig. 13 Illustration of the clustering behavior of the Q_β on $\rho_{\text{three blobs}}$ for $n = 966$ samples at time $t = 10$. The first three rows show the clustering behavior for $k = 2, 3, 4$, with each node colored according to which cluster it belongs. In the fourth row, we show the normalized k-means energy for each choice of k.



first three rows show the clustering behavior for $k = 2, 3, 4$, with each node colored according to which cluster it belongs. In the fourth row, we show the normalized k-means energy for each choice of k.

This simulation provides an example of a data distribution where there is no single "correct" choice of clustering for $k = 2$: a "good" clustering algorithm might seek to cut either the thin, high density rectangle on the bottom or the wide, low density rectangle on the top. For small values of $\beta \leqslant 0.7$, diffusion dominates, and the clusters are chosen based on the geometry of the data, preferring to cut the thin, high density rectangle. For large values of $\beta \geqslant 0.75$, density dominates, and the clustering prefers to cut the wide, low density rectangle. For intermediate values of β, there is a phase transition for which the clustering becomes unstable.

Acknowledgments The authors would like to thank Eric Carlen, Paul Atzberger, Anna Little, James Murphy, and Daniel McKenzie for helpful discussions. KC was supported by NSF DMS grant 1811012 and a Hellman Faculty Fellowship. NGT was supported by NSF-DMS grant 2005797. DS was supported by NSF grant DMS 1814991. NGT would also like to thank the IFDS at UW-Madison and NSF through TRIPODS grant 2023239 for their support. The authors gratefully acknowledge the support from the Simons Center for Theory of Computing, at which this work was completed.

References

1. L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
2. E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *J. Mach. Learn. Res.*, 17:Paper No. 43, 28, 2016.
3. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
4. M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *International Conference on Computational Learning Theory*, pages 486–500. Springer, 2005.
5. D. Burago, S. Ivanov, and Y. Kurylev. A graph discretization of the Laplace-Beltrami operator. *J. Spectr. Theory*, 4(4):675–714, 2014.
6. J. Calder and N. García Trillos. Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-nn graphs. *arXiv preprint arXiv:1910.13476*, 2019.
7. J. Calder, N. G. Trillos, and M. Lewicka. Lipschitz regularity of graph Laplacians on random data clouds, 2020.
8. M. A. Carreira-Perpiñán. Gaussian mean-shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.
9. M. A. Carreira-Perpiñán. Clustering methods based on kernel density estimators: mean-shift algorithms. In *Handbook of cluster analysis*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 383–417. CRC Press, Boca Raton, FL, 2016.
10. X. Cheng and H.-T. Wu. Convergence of graph Laplacian with KNN self-tuned kernels. *arXiv preprint arXiv:2011.01479*, 2020.
11. S.-N. Chow, L. Dieci, W. Li, and H. Zhou. Entropy dissipation semi-discretization schemes for Fokker-Planck equations. *J. Dynam. Differential Equations*, 31(2):765–792, 2019.
12. R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.
13. D. B. Dunson, H.-T. Wu, and N. Wu. Spectral convergence of graph Laplacian and heat kernel reconstruction in l^∞ from random samples, 2019.
14. A. Esposito, F. S. Patacchini, A. Schlichting, and D. Slepcev. Nonlocal-interaction equation on graphs: gradient flow structure and continuum limit. *Arch. Ration. Mech. Anal.*, 240(2):699–760, 2021.

15. K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.
16. N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*, pages 1–61, 2019.
17. N. García Trillos, F. Hoffmann, and B. Hosseini. Geometric structure of graph Laplacian embeddings. *Journal of Machine Learning Research*, 22(63):1–55, 2021.
18. N. García Trillos and D. Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239–281, 2018.
19. E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.
20. M. Hein, J.-Y. Audibert, and U. v. Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(6), 2007.
21. M. Hein, J.-Y. Audibert, and U. Von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005.
22. V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, Oct 2004.
23. J. D. Hunter. Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.*, 9(3):90–95, 2007.
24. E. Jones, T. Oliphant, P. Peterson, et al. *SciPy: Open source scientific tools for Python*, 2001–. Available at http://www.scipy.org/.
25. W. L. G. Koontz, P. M. Narendra, and K. Fukunaga. A graph-theoretic approach to nonparametric cluster analysis. *IEEE Transactions on Computers*, (9):936–944, 1976.
26. A. Little, M. Maggioni, and J. M. Murphy. Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *Journal of Machine Learning Research*, 21(6):1–66, 2020.
27. A. Little, D. McKenzie, and J. Murphy. Balancing geometry and density: Path distances on high-dimensional data, 2020.
28. J. Lu. Graph approximations to the Laplacian spectra. *arXiv: Differential Geometry*, 2019.
29. J. Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 8(261):2250–2292, 2011.
30. L. Michel. About small eigenvalues of the Witten Laplacian. *Pure and Applied Analysis*, 1(2):149–206, 2019.
31. B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006. Special Issue: Diffusion Maps and Wavelets.
32. A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856. MIT Press, 2001.
33. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
34. G. Schiebinger, M. J. Wainwright, and B. Yu. The geometry of kernelized spectral clustering. *The Annals of Statistics*, 43(2):819–846, 2015.
35. J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.
36. Z. Shi. Convergence of Laplacian spectra from random samples. arXiv preprint arXiv:1507.00151, 2015.
37. A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

38. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.
39. H. tieng Wu and N. Wu. When locally linear embedding hits boundary, 2018.
40. D. Ting, L. Huang, and M. Jordan. An analysis of the convergence of graph laplacians. *arXiv preprint arXiv:1101.5435*, 2011.
41. L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality reduction: A comparative review. Tilburg University Technical Report, TiCC-TR2009-005. 2009.
42. S. van der Walt, C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, 13(2):22–30, 2011.
43. R. Vaughn, T. Berry, and H. Antil. Diffusion maps for embedded manifolds with boundary with applications to PDEs, 2019.
44. A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *European conference on computer vision*, pages 705–718. Springer, 2008.
45. U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007.
46. U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.
47. E. Witten. Supersymmetry and Morse theory. *Journal of Differential Geometry*, 17(4):661–692, 1982.
48. C. L. Wormell and S. Reich. Spectral convergence of diffusion maps: improved error bounds and an alternative normalisation, 2020.

Random Batch Methods for Classical and Quantum Interacting Particle Systems and Statistical Samplings

Shi Jin and Lei Li

Abstract We review the Random Batch Methods (RBM) for interacting particle systems consisting of N-particles, with N being large. The computational cost of such systems is of $\mathcal{O}(N^2)$, which is prohibitively expensive. The RBM methods use small but random batches so the computational cost is reduced, per time step, to $\mathcal{O}(N)$. In this article we discuss these methods for both classical and quantum systems, the corresponding theory, and applications from molecular dynamics, statistical samplings, to agent-based models for collective behavior, and quantum Monte Carlo methods.

1 Introduction

Interacting particle systems arise in a variety of important phenomena in physical, social, and biological sciences. They usually take the form of Newton's second law that governs the interactions of N-particles under interacting forces that vary depending on different applications. Such systems are important in physics–from electrostatics to astrophysics, in chemistry and material sciences–such as molecular dynamics, in biological and social sciences–such as agent-based models in swarming [95, 16, 15, 21], chemotaxis [51, 8], flocking [19, 45, 2], synchronization [17, 44], and consensus [83]).

These interacting particle systems can be described in general by the first order systems

$$d\boldsymbol{r}_i = b(\boldsymbol{r}_i)\,dt + \alpha_N \sum_{j:j\neq i} K(\boldsymbol{r}_i - \boldsymbol{r}_j)\,dt + \sigma\, d\boldsymbol{W}_i, \quad i = 1, 2, \cdots, N, \tag{1.1}$$

or the second order systems

S. Jin (✉) · L. Li
School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai, P. R. China
e-mail: shijin-m@sjtu.edu.cn; leili2010@sjtu.edu.cn

N. Bellomo et al. (eds.), *Active Particles, Volume 3*, Modeling and Simulation in Science, Engineering and Technology, https://doi.org/10.1007/978-3-030-93302-9_5

$$\begin{aligned} d\boldsymbol{r}_i &= \boldsymbol{v}_i\, dt, \\ d\boldsymbol{v}_i &= \Big[b(\boldsymbol{r}_i) + \alpha_N \sum_{j: j\neq i} K(\boldsymbol{r}_i - \boldsymbol{r}_j) - \gamma \boldsymbol{v}_i \Big] dt + \sigma\, d\boldsymbol{W}_i. \end{aligned} \tag{1.2}$$

We use $i = 1, \cdots, N$ to denote the labels for the particles. We will loosely call $\boldsymbol{r}_i$ the "locations" or "positions," and $\boldsymbol{v}_i$ the velocities of the particles, though the specific meaning can be different in different applications. The function $K(\cdot)$ and $b(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ are the interaction kernel and some given external field respectively. The stochastic processes $\{\boldsymbol{W}^i\}_{i=1}^N$ are i.i.d. Wiener processes, or the standard Brownian motions. If $\gamma = \sigma = 0$ and $b = -\nabla V$ for some potential V, one has a Hamiltonian system in classical mechanics. For the molecules in the heat bath [62, 14], $\boldsymbol{r}_i$ and $\boldsymbol{v}_i$ are the physical positions and velocities, described by the underdamped Langevin equations, where σ and γ satisfy the so-called fluctuation-dissipation relation

$$\sigma = \sqrt{2\gamma/\beta}, \tag{1.3}$$

where β is the inverse of the temperature. (we assume all the quantities are scaled and hence dimensionless so that the Boltzmann constant is absent.) The first order system (1.1) can be viewed as the overdamped limit (when $\gamma \to \infty$ and the time rescaled as γt) of the second order systems (1.2). When the fluctuation-dissipation relation (1.3) is satisfied so that the diffusion coefficient in (1.1) is given by $\sqrt{2/\beta}$.

In the case $\alpha_N = \frac{1}{N-1}$, as $N \to \infty$, the dynamics of the so-called mean-field limit of (1.1) is given by the nonlinear Fokker-Planck equation

$$\partial_t \mu = -\nabla \cdot ((b(x) + K * \mu)\mu) + \frac{1}{2}\sigma^2 \Delta\mu, \tag{1.4}$$

where $\mu(dx) \in \mathbf{P}(\mathbb{R}^d)$. (the notation $\mathbf{P}(E)$ for a Polish space E means the set of all probability measures on E.) This means that the empirical measure defined by

$$\mu_N(t) := \frac{1}{N} \sum_{i=1}^N \delta(\cdot - \boldsymbol{r}_i(t)), \tag{1.5}$$

and the one-particle marginal distribution converges (in the weak topology) to the weak solution of the Eq. (1.4). See [79, 92, 37, 40, 52, 69] for some references about the mean-field limit. The regime $\alpha_N = c/N + o(1/N)$, $N \to \infty$ is thus naturally called the mean-field regime. Correspondingly, the mean-field limit of the second order system (1.2) is

$$\partial_t f = -\nabla_x \cdot (vf) - \nabla_v \cdot ((b(x) + K *_x f - \gamma v) f) + \frac{1}{2}\sigma^2 \Delta_v f, \tag{1.6}$$

where $f(dx, dv) \in \mathbf{P}(\mathbb{R}^d \times \mathbb{R}^d)$ and $*_x$ means that the convolution is performed only on the x variable.

If one directly discretizes (1.1) or (1.2), the computational cost per time step is $\mathcal{O}(N^2)$. This is undesired for large N. The Fast Multipole Method (FMM) [90] is able to reduce the complexity to $\mathcal{O}(N)$ for fast enough decaying interactions. However, the implementation of FMM is quite involved. A simple random algorithm, called the Random Batch Method (RBM), has been proposed in [54] to reduce the computation cost per time step from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, based on the simple "random mini-batch" idea. Such an idea is famous for its application in the so-called stochastic gradient descent (SGD) [89, 11, 13] for machine learning problems. The idea was also used for Markov Chain Monte Carlo methods like the stochastic gradient Langevin dynamics (SGLD) [98], and the computation of the mean-field flocking model [2, 16], motivated by Nanbu's algorithm of the Direct Simulation Monte Carlo method [10, 84, 5].

The key behind the "mini-batch" idea is to find some cheap unbiased random estimator using small subset of data/particles for the original quantity with the variance being controlled. Depending on the specific applications, the design can be different. For instances, the random grouping strategy was proposed in the RBM regarding general interacting particle systems in [54] (see also Sect. 2 and Lemma 2.1 below for details), while the importance sampling in the Fourier space was proposed for the Random Batch Ewald method for molecular dynamics in [59]. Compared with FMM, the accuracy of RBM is lower, but RBM is much simpler and is valid for more general potentials (e.g., the SVGD ODE [67]). The method converges due to the time average, and thus the convergence is like that in the Law of Large Numbers, but in time. For long-time behaviors, the method works for systems that own ergodicity and mixing properties, like systems in contact with heat bath and converging to equilibria. A key difference from SGD or SGLD is that the RBM algorithms proposed are aiming to approximate and grasp the dynamical properties of the systems as well, not just to find the optimizer or equilibrium distribution.

The rigorous analysis of RBM has been established for some cases and RBM has been shown to be asymptotic-preserving in the mean-field limit [54]. RBM for interacting particle systems has already been used or extended in various directions, from statistical sampling [67, 70, 57] to molecular dynamics [59, 68], control of synchronization [9, 65], and collective behavior of agent-based models [46, 43, 64]. RBM has been shown to converge for finite time interval if the interaction kernels are good enough [67, 54], and in particular, an error estimate *uniformly* in N was first obtained in [54]. A convergence result of RBM for N-body Schrödinger equation was established in [41].

The goal of this review is to introduce the basics of the RBM, the fundamental theory for the convergence and error estimates, and various applications.

2 The Random Batch Methods

In this section, we describe the RBM for general interacting particle systems introduced first in [54]. We use bold fonts (e.g., $\boldsymbol{r}_i$, $\boldsymbol{x}_i$, $\boldsymbol{v}_i$, $\boldsymbol{u}_i$) and capital letters (X_i, Y_i) to denote the quantities that are functions of time t associated with the particles, use usual letters like x_i, v_i to represent a point in the state space (often $\mathbb{R}^d$), and use letters like $\underline{x}$, $\underline{v}$ to represent quantities in the configurational space $\mathbb{R}^{Nd}$.

2.1 The RBM Algorithms

Let $T > 0$ be the simulation time, and choose a time step $\Delta t > 0$. Pick a batch size $2 \le p \ll N$ that divides N (RBM can also be applied if p does not divide N; we assume this only for convenience). Consider the discrete time grids $t_k := k\Delta t$, $k \in \mathbb{N}$. For each subinterval $[t_{k-1}, t_k)$, the method has two substeps: (1) at t_{k-1}, divide the N particles into $n := N/p$ groups (batches) randomly; (2) let the particles evolve with interaction only inside the batches.

The above procedure, when applied to the second order system (1.2), leads to Algorithm 1. The version for first order systems is similar.

RBM requires the random division, and the elements in different batches are different. This is in fact the sampling without replacement. If one allows replacement, one has the following version of RBM 2, which is simpler to implement. In this version, for one iteration of k, some particles may not be updated while some maybe drawn more than once. However, the method is expected to be correct statistically.

Algorithm 1 (RBM for (1.2))

```
for $m$ in $1 : [T/\Delta t]$ do
    Divide $\{1, 2, \dots, N = pn\}$ into $n$ batches $\mathcal{C}_q$, $1 \le q \le n$ randomly.
    for each batch $\mathcal{C}_q$ do
        Update $\boldsymbol{r}_i, \boldsymbol{v}_i$ ($i \in \mathcal{C}_q$) by solving for $t \in [t_{m-1}, t_m)$ the following
```

$$
\begin{aligned}
d\boldsymbol{r}_i &= \boldsymbol{v}_i\, dt,\\
d\boldsymbol{v}_i &= \Big[b(\boldsymbol{r}_i) + \frac{\alpha_N(N-1)}{p-1}\sum_{j\in\mathcal{C}_q, j\neq i} K(\boldsymbol{r}_i - \boldsymbol{r}_j) - \gamma \boldsymbol{v}_i\Big] dt + \sigma\, d\boldsymbol{W}_i.
\end{aligned}
\tag{2.1}
$$

```
    end for
end for
```

Algorithm 2 (RBM-r)

```
for $m$ in $1 : [T/\Delta t]$ do
    for $k$ from 1 to $N/p$ do
        Pick a set $\mathcal{C}_k$ of size $p$ randomly with replacement.
        Update $\boldsymbol{r}_i$'s ($i \in \mathcal{C}_k$) by solving the following SDE for time $\Delta t$.
            $$\begin{cases} d\boldsymbol{x}_i = \boldsymbol{u}_i\, dt, \\ d\boldsymbol{u}_i = \Big[b(\boldsymbol{x}_i) + \frac{\alpha_N(N-1)}{p-1} \sum_{j\in\mathcal{C}_k, j\neq i} K(\boldsymbol{x}_i - \boldsymbol{x}_j) - \gamma \boldsymbol{u}_i\Big] dt + \sigma\, d\boldsymbol{W}_i. \\ \boldsymbol{x}_i(0) = \boldsymbol{r}_i, \quad \boldsymbol{u}_i(0) = \boldsymbol{v}_i, \end{cases} \tag{2.2}$$
        i.e., solve (2.2) with initial values $\boldsymbol{x}_i(0) = \boldsymbol{r}_i, \boldsymbol{u}_i(0) = \boldsymbol{v}_i$, and set $\boldsymbol{r}_i \leftarrow \boldsymbol{x}_i(\Delta t)$, $\boldsymbol{v}_i \leftarrow \boldsymbol{u}_i(\Delta t)$.
    end for
end for
```

We now discuss the computational cost. Note that random division into n batches of equal size can be implemented using random permutation, which can be realized in $\mathcal{O}(N)$ operations by Durstenfeld's modern revision of Fisher-Yates shuffle algorithm [28] (in MATLAB, one can use "randperm(N)"). After the permutation, one takes the first p elements to be in the first batch, the second p elements to be in the second batch, etc. The ODE solver per particle per time step (2.2) requires merely $\mathcal{O}(p)$ operations, thus for all particles, each time step costs only $\mathcal{O}(pN)$. Since $p \ll N$ the overall cost per time step is significantly reduced from $\mathcal{O}(N^2)$.

However, one might encounter the issue of having to use a much smaller time step, which could be of $\mathcal{O}(N)$ times smaller, in the RBM implementation. For RBM to really gain significant efficiency, one needs Δt to be *independent* of N. This is justified by an error analysis to be presented in the next subsection.

2.2 Convergence Analysis

In this subsection, we present the convergence results of RBM for the second order systems (1.2) in the mean-field regime (i.e., $\alpha_N = 1/(N-1)$), which was given in [56]. We remark that the proof relies on the underlying contraction property of the second order systems under certain conditions [78, 30]. Due to the degeneracy of the noise terms, the contraction should be proved by suitably chosen variables and Lyapunov functions, and we refer the readers to [56] for more details.

Denote $(\tilde{\boldsymbol{r}}_i, \tilde{\boldsymbol{v}}_i)$ the solutions to the random batch process (2.1) with the Brownian motion used being $\tilde{\boldsymbol{W}}_i$. Consider the synchronization coupling as in [54, 55]:

$$\boldsymbol{r}_i(0) = \tilde{\boldsymbol{r}}_i(0) \sim \mu_0, \quad \boldsymbol{W}_i = \tilde{\boldsymbol{W}}_i. \tag{2.3}$$

Let $\mathcal{C}_q^{(k)}$ $(1 \le q \le n)$ be the batches at t_k, and define

$$\mathcal{C}^{(k)} := \{\mathcal{C}_1^{(k)}, \cdots, \mathcal{C}_n^{(k)}\}, \tag{2.4}$$

to be the random division of batches at t_k. According to the Kolmogorov extension theorem [27], there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the random variables $\{\boldsymbol{r}_0^i, W^i, \mathcal{C}^{(k)} : 1 \le i \le N, k \ge 0\}$ are all defined on this probability space and are independent. Let $\mathbb{E}$ denote the integration on Ω with respect to the probability measure $\mathbb{P}$, and consider the $L^2(\cdot)$ norm of a random variable

$$\|\zeta\| = \sqrt{\mathbb{E}|\zeta|^2}. \tag{2.5}$$

For finite time interval, the convergence of RBM is as follows.

Theorem 2.1 *Let $b(\cdot)$ be Lipschitz continuous, and assume that $|\nabla^2 b|$ has polynomial growth, and the interaction kernel K is Lipschitz continuous. Then,*

$$\sup_{t\in[0,T]} \sqrt{\mathbb{E}|\tilde{\boldsymbol{r}}_i(t) - \boldsymbol{r}_i(t)|^2 + \mathbb{E}|\tilde{\boldsymbol{v}}_i(t) - \boldsymbol{v}_i(t)|^2} \le C(T)\sqrt{\frac{\Delta t}{p-1} + \Delta t^2}, \tag{2.6}$$

where $C(T)$ is independent of N.

Often the long-time error estimates are important since one could use RBM as a sampling method for the invariant measure of (1.2) (see Sect. 5). For this we need some additional contraction assumptions:

Assumption 2.1 $b = -\nabla V$ for some $V \in C^2(\mathbb{R}^d)$ that is bounded from below (i.e., $\inf_x V(x) > -\infty$), and there exist $\lambda_M \ge \lambda_m > 0$ such that the eigenvalues of $H := \nabla^2 V$ satisfy

$$\lambda_m \le \lambda_i(x) \le \lambda_M,\ \forall\, 1 \le i \le d, x \in \mathbb{R}^d.$$

The interaction kernel K is bounded and Lipschitz continuous. Moreover, the friction γ and the Lipschitz constant L of $K(\cdot)$ satisfy

$$\gamma > \sqrt{\lambda_M + 2L},\ \ \lambda_m > 2L. \tag{2.7}$$

Then the following uniform strong convergence estimate holds:

Theorem 2.2 *Under Assumption 2.1 and the coupling (2.3), the solutions to (1.2) and (2.1) satisfy*

$$\sup_{t\ge 0} \sqrt{\mathbb{E}|\tilde{\boldsymbol{r}}_i(t) - \boldsymbol{r}_i(t)|^2 + \mathbb{E}|\tilde{\boldsymbol{v}}_i(t) - \boldsymbol{v}_i(t)|^2} \le C\sqrt{\frac{\Delta t}{p-1} + \Delta t^2}, \tag{2.8}$$

where the constant C does not depend on p and N.

Clearly, these error estimates imply that the RBM can also grasp the dynamical properties. The error estimates above are consequence of some intuitive results, which we summarize here (see [54]).

For given $\underline{x} := (x_1, \ldots, x_N) \in \mathbb{R}^{Nd}$, introduce the error of the interacting force for the ith particle.

$$\chi_i(\underline{x}) := \frac{1}{p-1}\sum_{j\in\mathcal{C}} K(x_i - x_j) - \frac{1}{N-1}\sum_{j:j\neq i} K(x_i - x_j). \tag{2.9}$$

Here, $\mathcal{C}$ is the random batch that contains i in a random division of the batches.

Lemma 2.1 *Consider a configuration $\underline{x}$ that is independent of the random division. Then,*

$$\mathbb{E}\chi_i(\underline{x}) = 0. \tag{2.10}$$

Moreover, the (scalar) variance is given by

$$\mathrm{Var}(\chi_i(\underline{x})) = \left(\frac{1}{p-1} - \frac{1}{N-1}\right)\Lambda_i(\underline{x}), \tag{2.11}$$

where

$$\Lambda_i(\underline{x}) := \frac{1}{N-2}\sum_{j:j\neq i}\left|K(x_i - x_j) - \frac{1}{N-1}\sum_{\ell:\ell\neq i} K(x_i - x_\ell)\right|^2. \tag{2.12}$$

Lemma 2.1 in fact lays the foundation of the convergence of RBM-type algorithms. The first claim implies that the random estimates of the interacting forces are unbiased in the sense that the expectation is zero. This gives the consistency–in expected value–of the RBM approximation, although each random batch approximation $\frac{1}{p-1}\sum_{j\in\mathcal{C}} K(x_i - x_j)$ to the true interacting force $\frac{1}{N-1}\sum_{j:j\neq i} K(x_i - x_j)$ gives an $\mathcal{O}(1)$ error (which is clear from $\sqrt{\mathrm{Var}(\chi_i(\underline{x}))} = \mathcal{O}(1)$). Being a Monte Carlo like methods, the boundedness of the variance ensures the stability of the RBM as can be seen in the proof [56, 55]. The intrinsic mechanism why such type of methods work is the independent resampling in later time steps, and due to some averaging effect in time these $\mathcal{O}(1)$ errors become small. This Law of Large Numbers type feature *in time* guarantees the convergence of RBMs (as indicated by the error bound $\sqrt{|\mathrm{Var}(\chi)|\tau} \sim \sqrt{|\mathrm{Var}(\chi)|/N_T}$ in Theorems 2.1 and 2.2).

As another remark, the nonzero variance of the RBM approximation gives some effective noise into the system. This could bring in some "numerical heating" effects when RBM is applied for some interacting particle systems. When the system has some dissipation, or in contact with a heat bath as in Sect. 4 , RBM approximation can be valid for long time and can capture the equilibrium.

In both Theorems 2.1 and 2.2, the error bound is independent of N so that the time step can be chosen independent of N for a fixed accuracy tolerance ε. Hence, for each time step, the cost of RBM is $\mathcal{O}(1/N)$ of that for direct simulation, but it does not need to take $\mathcal{O}(N)$ times longer to finish the computation. Such convergence results were first established for first order systems (1.1) [54] and then extended to disparate mass cases [55]. The weak convergence has also been discussed in [55].

2.3 An Illustrating Example: Wealth Evolution

To illustrate the algorithms, we consider the model proposed by Degond et. al. [22] for the evolution of N market agents with two attributes: the economic configuration X_i and its wealth Y_i.

$$\begin{aligned} \dot{X}_i &= V(X_i, Y_i), \\ dY_i &= -\frac{1}{N-1}\sum_{k:k\neq i} \xi_{ik}\Psi(|X_i - X_k|)\partial_y\phi(Y_i - Y_k)\,dt + \sqrt{2D}Y_i dW_i. \end{aligned} \tag{2.13}$$

The first equation describes the evolution of the economic configuration, which is driven by the local Nash equilibrium and it is related to mean-field games [66]. The second equation describes the evolution of the wealth. The quantity $\sqrt{2D}$ is the volatility. The function ϕ is the trading interaction potential, while $\xi_{ik}\Psi(|X^i - X^k|)$ is the trading frequency. This model is an interacting particle systems with long-range interactions and multiplicative noise, for which we will apply the RBM method. We also point out that the RBM version of (2.13) can be viewed as a *new model* in which one agent may only trade with a small number of random agents during a short time in the real world.

For numerical experiments, [54] considers the homogeneous case when the wealth dynamics is independent of the position in the economic configuration space. Then, the dynamics of the wealth is reduced to the following

$$dY_i = -\frac{\kappa}{N-1}\sum_{k:k\neq i} \partial_y\phi(Y_i - Y_k)\,dt + \sqrt{2D}Y_i dW_i. \tag{2.14}$$

The corresponding mean-field dynamics has an equilibrium distribution given by

$$\rho_\infty(y) \propto \exp\left(-\frac{\alpha(y)}{D}\right),$$

where α satisfies

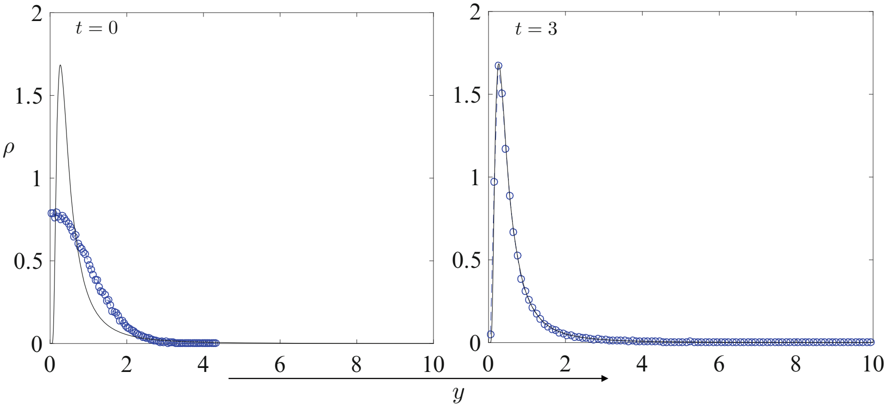


Fig. 1 Wealth distribution obtained by RBM compared with the reference curve

$$\partial_y \alpha(y) = -\frac{1}{y^2}F(y) + \frac{2D}{y}.$$

In Fig. 1, the empirical distribution of the wealth obtained by RBM for the case $\phi(y) = \frac{1}{2}y^2$ is compared to the reference curve (an inverse Gamma distribution), which is

$$\rho_\infty(y) = \frac{(\kappa\eta/D)^{\kappa/D+1}}{\Gamma(\kappa/D+1)} y^{-(2+\kappa/D)} \exp\left(-\frac{\kappa\eta}{Dy}\right) 1_{y>0}, \quad \eta = \frac{\sqrt{2}}{\sqrt{\pi}}.$$

Clearly, the distribution obtained by RBM agrees perfectly with the expected wealth distribution at $t = 3$ already.

This example has two distinguished features: long range and multiplicative noises so that it does not fit the assumptions of the convergence results presented in Sect. 2.2, which were established for regular interacting potentials K and additive noises. As shown by this and more examples in [54], and those in later sections, the RBM algorithms are applicable to much broader classes of interacting particle systems, including long-range, singular (like the Lennard-Jones and Coulomb) potentials (see Sect. 4 below), and with multiplicative noise.

3 The Mean-Field Limit

It is known that the N-particle system (1.1) with $\alpha_N = 1/(N-1)$ has the mean-field limit given by the Fokker-Planck equation (1.4). Namely, the empirical measure or the one-particle marginal distribution of the particle system (1.1) is close, in Wasserstein distance, to μ in (1.4). Thus, when N is large, one may use the RBM

Algorithm 3 (RBM for first order systems)

```
for $k$ in $1 : [T/\Delta t]$ do
    Divide $\{1, 2, \dots, N\}$ into $n = N/p$ batches randomly.
    for each batch $\mathcal{C}_q$ do
        Update $\boldsymbol{r}_i$'s ($i \in \mathcal{C}_q$) by solving the following SDE with $t \in [t_{k-1}, t_k)$.
            $$d\boldsymbol{r}_i = b(\boldsymbol{r}_i)dt + \frac{1}{p-1}\sum_{j\in\mathcal{C}_q, j\neq i} K(\boldsymbol{r}_i - \boldsymbol{r}_j)dt + \sigma\, d\boldsymbol{W}^i. \quad (3.1)$$
    end for
end for
```

Algorithm 4 (Mean-field dynamics of RBM (3.1))

```
$\tilde{\mu}(\cdot, t_0) = \mu_0$.
for $k \ge 0$ do
    Let $\rho^{(p)}(\cdots, 0) = \tilde{\mu}(\cdot, t_k)^{\otimes p}$ be a probability measure on $(\mathbb{R}^d)^p \cong \mathbb{R}^{pd}$.
    Evolve the measure $\rho^{(p)}$ to find $\rho^{(p)}(\cdots, \Delta t)$ by the following Fokker-Planck equation:
        $$\partial_t \rho^{(p)} = -\sum_{i=1}^{p} \nabla_{x_i} \cdot \left(\Big[b(x_i) + \frac{1}{p-1}\sum_{j=1, j\neq i}^{p} K(x_i - x_j)\Big]\rho^{(p)}\right) + \frac{1}{2}\sigma^2 \sum_{i=1}^{p} \Delta_{x_i}\rho^{(p)}. \quad (3.2)$$
    Set
        $$\tilde{\mu}(\cdot, t_{k+1}) := \int_{(\mathbb{R}^d)^{(p-1)}} \rho^{(p)}(\cdot, dy_2, \cdots, dy_p, \Delta t). \quad (3.3)$$
end for
```

as a numerical (particle method) for (1.4). Indeed, since the error bounds obtained in the previous section are independent of N, one could hope that when $N \to \infty$, the one-particle marginal distribution of the RBM should be close to μ. To justify this, one first needs to derive the mean-field limit of the RBM, for fixed Δt, then compare it with (1.4). In addition, the RBM could be viewed as a random model for the underlying physics, hence it is also natural to ask what its mean-field limit is.

Consider the RBM for the first order system (1.1) with $\alpha_N = 1/(N-1)$, shown in Algorithm 3. The mean-field limit was derived and proved in [53]. We summarize the results in this section.

Intuitively, when $N \gg 1$, the probability that two chosen particles are correlated is very small. Hence, in the $N \to \infty$ limit, two chosen particles will be independent with probability 1. Due to the exchangeability, the marginal distributions of the particles will be identical. Based on this observation, the following mean-field limit for RBM can be obtained for the one-particle distribution:

The dynamics in Algorithm 4 naturally gives a nonlinear operator $\mathcal{G}_\infty$: $\mathbf{P}(\mathbb{R}^d) \to \mathbf{P}(\mathbb{R}^d)$ as

$$\tilde{\mu}(\cdot, t_{k+1}) =: \mathcal{G}_\infty(\tilde{\mu}(\cdot, t_k)). \tag{3.4}$$

Corresponding to this is the following SDE system for $t \in [t_k, t_{k+1})$

$$d\boldsymbol{x}_i = b(\boldsymbol{x}_i)\,dt + \frac{1}{p-1}\sum_{j=1, j\neq i}^{p} K(\boldsymbol{x}_i - \boldsymbol{x}_j)\,dt + \sigma\, d\boldsymbol{W}_i, \ \ i = 1, \cdots, p, \tag{3.5}$$

with $\{\boldsymbol{x}_i(t_k)\}$ drawn i.i.d from $\tilde{\mu}(\cdot, t_k)$. Then, $\tilde{\mu}(\cdot, t_{k+1}) = \mathscr{L}(\boldsymbol{x}_1(t_{k+1}^-))$, the law of $\boldsymbol{x}_1(t_{k+1}^-)$. Note that all $\boldsymbol{x}_i$ have the same distribution for any $t_k \le t < t_{k+1}$. Without loss of generality, we will impose $\boldsymbol{x}_1(t_k^-) = \boldsymbol{x}_1(t_k^+)$. For other particles $i \neq 1$, $\boldsymbol{x}_i(t)$ in $[t_{k-1}, t_k)$ and $[t_k, t_{k+1})$ are independent and they are not continuous at t_k. In fact, in the $N \to \infty$ limit, $\boldsymbol{x}_i, i \neq 1$ at different subintervals correspond to different particles that interact with particle 1 as in Algorithm 3.

Hence, in the mean-field limit of RBM, one starts with a chaotic configuration,[1] the p particles evolve by interacting with each other. Then, at the starting point of the next time interval, one imposes the chaos condition so that the particles are independent again.

In [53], this intuition has been justified rigorously for finitely many steps under the following assumptions.

Assumption 3.1 The moments of the initial data are finite:

$$\int_{\mathbb{R}^d} |x|^q \mu_0(dx) < \infty, \ \forall q \in [2, \infty). \tag{3.6}$$

Assumption 3.2 Assume $b(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ and $K(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ satisfy the following conditions.

- $b(\cdot)$ is one-sided Lipschitz:

$$(z_1 - z_2)\cdot(b(z_1) - b(z_2)) \le \beta |z_1 - z_2|^2 \tag{3.7}$$

  for some constant β;
- K is Lipschitz continuous

$$|K(z_1) - K(z_2)| \le L|z_1 - z_2|.$$

[1] By "chaotic configuration," we mean that there exists a one-particle distribution f such that for any j, the j-marginal distribution is given by $\mu^{(j)} = f^{\otimes j}$. Such independence in a configuration is then loosely called "chaos." If the j-marginal distribution is more close to $f^{\otimes j}$ for some f, we loosely say "there is more chaos."

One marginal of N particle system Nonlinear Fokker-Planck

$$\begin{tikzcd}
\mu_N^{(1)} \arrow[r, "N \to \infty"] & \mu = \varrho\, dx \quad \mu(\cdot,t) = \mathcal{S}(t)(\mu_0) \\
\tilde{\mu}_N^{(1)} \arrow[u, "\Delta t \to 0"] \arrow[r, "N \to \infty"'] & \tilde{\mu} \arrow[u, "\Delta t \to 0"']
\end{tikzcd}$$

One marginal of RBM Mean field limit of RBM

$\tilde{\mu}_N^{(1)}(\cdot, t_k) = \mathcal{G}_N^{(k)}(\mu_0)$ $\tilde{\mu}(\cdot, t_k) = \mathcal{G}_\infty^k(\mu_0)$

Fig. 2 Illustration of the various operators and the asymptotic limits

Corresponding to the operator (3.4), one may define the operator $\mathcal{G}_N^k : \mathbf{P}(\mathbb{R}^d) \to \mathbf{P}(\mathbb{R}^d)$ for RBM with N particles as follows. Let $\boldsymbol{r}_i(0)$'s be i.i.d drawn from μ_0, and consider (3.1). Define

$$\mathcal{G}_N^k(\mu_0) := \mathscr{L}(\boldsymbol{r}_1(t_k)). \tag{3.8}$$

Recall that $\mathscr{L}(\boldsymbol{r}_1)$ denotes the law of $\boldsymbol{r}_1$, thus the one-particle marginal distribution. Conditioning on a specific sequence of random batches, the particles are not exchangeable. However, when one considers the mixture of all possible sequences of random batches, the laws of the particles $\boldsymbol{r}_i(t_k)$ $(1 \le i \le N)$ are identical. In Fig. 2, we illustrate these definitions and various limits. With these setup introduced, we may state the first main result in [53] as follows:

Theorem 3.1 *Under assumptions 3.1 and 3.2, for any fixed k, it holds for any* $q \in [1, \infty)$ *that*

$$\lim_{N\to\infty} W_q(\mathcal{G}_\infty^k(\mu_0), \mathcal{G}_N^k(\mu_0)) = 0. \tag{3.9}$$

Here, W_q is the Wasserstein-q distance [91]:

$$W_q(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^q d\gamma \right)^{1/q}, \tag{3.10}$$

where $\Pi(\mu, \nu)$ is the set of "transport plans," i.e., a joint measure on $\mathbb{R}^d \times \mathbb{R}^d$ such that the marginal measures corresponding to x and y are μ and ν respectively.

The next questions are whether the one-particle marginal distribution $\mu_N^{(1)} := \mathscr{L}(\boldsymbol{r}_1)$ of the RBM converges to μ. Denote the solution operator to (1.4) by $\mathcal{S}$:

$$\mathcal{S}(\Delta)\mu(t_1) := \mu(t_1 + \Delta), \ \forall t_1 \ge 0, \Delta \ge 0. \tag{3.11}$$

Clearly, $\{\mathcal{S}(t) : t \ge 0\}$ is a nonlinear semigroup.

We make more technical assumptions here.

Assumption 3.3 The measure μ_0 has a density ϱ_0 that is smooth with finite moments $\int_{\mathbb{R}^d} |x|^q \varrho_0 \, dx < \infty$, $\forall q \geq 1$, and the entropy is finite

$$H(\mu_0) := \int_{\mathbb{R}^d} \varrho_0 \log \varrho_0 \, dx < \infty. \tag{3.12}$$

If $\varrho_0(x) = 0$ at some point x, one defines $\varrho_0(x) \log \varrho_0(x) = 0$. We also introduce the following assumption on the growth rate of derivatives of b and K, which will be used below.

Assumption 3.4 The function b and its derivatives have polynomial growth. The derivatives of K with order at least 2 (i.e., $D^\alpha K$ with $|\alpha| \geq 2$) have polynomial growth.

Based on these conditions, it can be shown that μ has a density $\varrho(\cdot, t)$. For convenience, we will not distinguish μ from its density ϱ. Sometimes, one may also assume the strong confinement condition:

Assumption 3.5 The fields $b(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ and $K(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ are smooth. Moreover, $b(\cdot)$ is strongly confining:

$$(z_1 - z_2) \cdot (b(z_1) - b(z_2)) \leq -r|z_1 - z_2|^2 \tag{3.13}$$

for some constant $r > 0$, and K is Lipschitz continuous $|K(z_1) - K(z_2)| \leq L|z_1 - z_2|$. The parameters r, L satisfy

$$r > 2L. \tag{3.14}$$

With the assumptions stated, we can state the second main result in [53].

Theorem 3.2 *Suppose Assumptions 3.2, 3.3 and 3.4 hold. Then,*

$$\sup_{n : n\Delta t \leq T} W_1(\mathcal{G}_\infty^n(\varrho_0), \varrho(n\Delta t)) \leq C(T)\Delta t. \tag{3.15}$$

If Assumption 3.5 is assumed in place of Assumption 3.2 and also $\sigma > 0$, then

$$\sup_{n \geq 0} W_1(\mathcal{G}_\infty^n(\varrho_0), \varrho(n\Delta t)) \leq C\Delta t. \tag{3.16}$$

These theorems show that the dynamics given by $\mathcal{G}_\infty$ can approximate that of the nonlinear Fokker-Planck equation (1.4), with the W_1 distance to be of $\mathcal{O}(\Delta t)$. Thus, the two limits $\lim_{N\to\infty}$ and $\lim_{\Delta t \to 0}$ commute, as shown in Fig. 2.

4 Molecular Dynamics

Molecular dynamics (MD) refers to computer simulation of atoms and molecules and is among the most popular numerical methods to understand the dynamical and equilibrium properties of many-body particle systems in many areas such as chemical physics, soft materials and biophysics [18, 34, 33]. In this section, we discuss the relevant issues and the applications of RBM and its modifications.

Consider N "molecules" with masses m_i's (each might be a model for a real molecule or a numerical molecule that is a packet of many real molecules) that interact with each other. The equations of motion are given by

$$\begin{aligned} d\boldsymbol{r}_i &= \boldsymbol{v}_i\, dt, \\ m_i d\boldsymbol{v}_i &= \Big[-\sum_{j:j\neq i} \nabla\phi_{ij}(\boldsymbol{r}_i - \boldsymbol{r}_j)\Big]\, dt + d\boldsymbol{\xi}_i. \end{aligned} \tag{4.1}$$

Here, $\phi_{ij}(\cdot)$ is the interaction potential and $d\boldsymbol{\xi}_i$ means some other possible terms that change the momentum,. Typical examples of the potential include the Coulomb potentials

$$\phi_{ij}(x) = \frac{q_i q_j}{r},$$

where q_i is the charge for the ith particle and $r = |x|$, and the Lennard-Jones potential

$$\phi_{ij}(x) = 4\left(\frac{1}{r^{12}} - \frac{1}{r^6}\right).$$

Between ions, both types of potential exist and between charge-neutral molecules, the Lennard-Jones potential might be the main force (the Lennard-Jones interaction intrinsically also arises from the interactions between charges, so these two types are in fact both electromagnetic forces) [34, 33]. To model the solids or fluids with large volume, one often uses a box with length L, equipped with the periodic conditions for the simulations. Below, we will assume $\phi_{ij} \equiv \phi$ independent of i, j for the convenience of discussions except explicitly stated otherwise.

To model the interaction between the molecules with the heat bath, one may consider some thermostats so that the temperature of the system can be controlled at a given value. The thermostats are especially good for RBM approximations as the effective noise introduced by RBM approximation can be damped by the thermostats, reducing the "numerical heating" effects [59]. Typical thermostats include the Andersen thermostat, the Langevin thermostat and the Nosé-Hoover thermostat [34]. In the Andersen thermostat [34, section 6.1.1], one does the simulation for $d\boldsymbol{\xi}_i = 0$ between two time steps, but a particle can collide with the heat bath at each discrete time. Specifically, assume the collision frequency is ν, so in a duration of time $t \ll 1$ the chance that a collision has happened is given by the exponential distribution

$$1 - \exp(-\nu t) \approx \nu t, \quad t \ll 1.$$

If a collision happens, the new velocity is then sampled from the Maxwellian distribution with temperature T (i.e., the normal distribution $\mathcal{N}(0, T)$). In the underdamped Langevin dynamics, one chooses

$$d\boldsymbol{\xi}_i = -\gamma \boldsymbol{v}_i \, dt + \sqrt{\frac{2\gamma}{\beta}} \, d\boldsymbol{W}_i,$$

so that the "fluctuation-dissipation relation" is satisfied and the system will evolve to the equilibrium with the correct temperature $T = \beta^{-1}$. It is well-known that the invariant measure of such systems is given by the Gibbs distribution [72]

$$\pi(\underline{x}, \underline{v}) \propto \exp\left(-\beta(\frac{1}{2}\sum_{i=1}^{N} |v_i|^2 + U(\underline{x}))\right), U(\underline{x}) = \frac{1}{2}\sum_{i,j:i\neq j} \phi(x_i - x_j),$$

where $\underline{x} = (x_1, \cdots, x_N) \in \mathbb{R}^{Nd}$ and $\underline{v} = (v_1, \cdots, v_N) \in \mathbb{R}^{Nd}$. The Nosé-Hoover thermostat uses a Hamiltonian for an extended system of N particles plus an additional coordinate s [85, 50]:

$$\mathcal{H}_{\mathrm{NH}} = \sum_{i=1}^{N} \frac{|\tilde{\boldsymbol{p}}_i|^2}{2m_i s^2} + U(\{\boldsymbol{r}_i\}) + \frac{p_s^2}{2Q} + L\frac{\ln s}{\beta}.$$

Here, $\tilde{\boldsymbol{p}}_i$ is the momentum of the ith particle. The microcanonical ensemble corresponding to this Hamiltonian reduces to the canonical ensemble for the real variables $\boldsymbol{p}_i = \tilde{\boldsymbol{p}}_i/s$. Hence, one may run the following deterministic ODEs, which are the Hamiltonian ODEs with Hamiltonian $\mathcal{H}_{\mathrm{NH}}$ in terms of the so-called real variables,

$$\begin{aligned}
\dot{\boldsymbol{r}}_i &= \boldsymbol{p}_i, \\
\dot{\boldsymbol{p}}_i &= -\nabla_{\boldsymbol{r}_i} U - \xi \boldsymbol{p}_i, \\
\dot{\xi} &= \frac{1}{Q}\left(\sum_{i=1} \frac{|\boldsymbol{p}_i|^2}{m_i} - \frac{3N}{\beta}\right).
\end{aligned}$$

The time average of the desired quantities will be the correct canonical ensemble average. As one can see, when the temperature of the system is different from T, the extra term $-\xi \boldsymbol{p}_i$ will drive the system back to temperature T, thus it may give better behaviors for controlling the temperature.

4.1 RBM with Kernel Splitting

In molecular dynamics simulation, the interaction force kernel

$$K(x) := -\nabla\phi(x),\ \ x \in \mathbb{R}^d,$$

is often singular at $x = 0$. Hence, the direct application of RBM could lead to poor results. The reason is that some particles in different batches may get too close after a time step and thus the interaction between them becomes large if they happen to be in the same batch in the next time interval, resulting in numerical instability. To resolve this issue, one can adopt the splitting strategy in [77, 49] and decompose the interacting force K into two parts:

$$K(x) = K_1(x) + K_2(x). \tag{4.2}$$

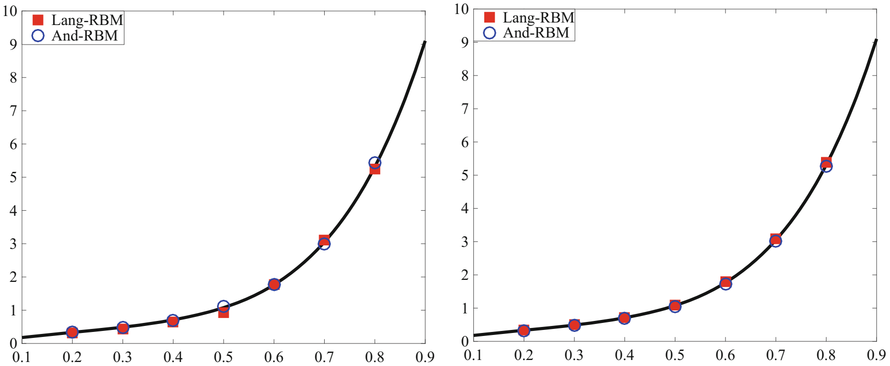
Here, K_1 has short range that vanishes for $|x| \ge r_0$ where r_0 is a certain cutoff chosen to be comparable to the mean distance of the particles. $K_2(x)$ is a bounded smooth function. One then applies RBM to the K_2 part only. The resulting method is shown in Algorithm 5. Now, K_1 is of short range so that one only considers the close neighbors to compute the summation in K_1 for each given i, and the resulting cost is of $\mathcal{O}(1)$ using data structures like Cell List [34, Appendix F]. Since K_2 is bounded, RBM can be applied well due to the boundedness of variance, without introducing too much error. Hence, the cost per time step is again $\mathcal{O}(N)$. For practical applications, one places the initial positions of the molecules on the grid of a lattice, and the repulsive force K_1 will forbid the particles from getting too close so that the system is not too stiff.

Algorithm 5 RBM with splitting for (1.2)

```
Split $K =: K_1 + K_2$, where $K_1$ has short range, while $K_2$ has long range but is smooth.
for $m$ in $1 : [T/\Delta t]$ do
    Divide $\{1, 2, \ldots, N = pn\}$ into $n$ batches randomly.
    for each batch $\mathcal{C}_q$ do
        Update $(\boldsymbol{r}_i, \boldsymbol{v}_i)$'s $(i \in \mathcal{C}_q)$ by solving for $t \in [t_{m-1}, t_m)$
```

$$\begin{aligned} d\boldsymbol{r}_i &= \boldsymbol{v}_i\,dt,\\ d\boldsymbol{v}_i &= \Big[b(\boldsymbol{r}_i) + \alpha_N \sum_{j:j\neq i} K_1(\boldsymbol{r}_i - \boldsymbol{r}_j) - \gamma \boldsymbol{v}_i\Big]dt\\ &\quad + \frac{\alpha_N(N-1)}{p-1} \sum_{j\in\mathcal{C}_q, j\neq i} K_2(\boldsymbol{r}_i - \boldsymbol{r}_j)\,dt + \sigma\,d\boldsymbol{W}_i. \end{aligned} \tag{4.3}$$

```
    end for
end for
```



(a)

(b)

Fig. 3 The pressure obtained by Andersen-RBM and Langevin-RBM for Lenard-Jones fluid with $N = 500$: the blue circles are those by Andersen-RBM while the red squares are by Langevin-RBM. The reference curves (black solid line) are the fitting curves in [60]. (**a**) $\nu = \gamma = 10$, $\Delta t_k = 0.001/\log(k + 1)$. (**b**) $\nu = \gamma = 50$, $\Delta t = 0.001$

Using this splitting strategy, one may apply RBM to the MD simulations with different thermostats. In Fig. 3, we show the numerical results from [56] for a Lennard-Jones fluid with temperature $\beta^{-1} = 2$ and the length of box $L = (N/\rho)^{1/3}$ for a given density ρ. The results are obtained using the Andersen thermostat and the Langevin thermostat respectively, with the splitting and RBM strategy, for particle number $N = 500$. In the first figure, the decreasing step sizes $\Delta t_k = 0.001/\log(k + 1)$ are taken to reduce the numerical heating effect brought by RBM when the collision coefficient are not so big ($\nu = \gamma = 10$). The results show that RBM with splitting strategy can work reasonably well for the Lennard-Jones fluid in the considered regime.

4.2 *Random Batch Ewald: An Importance Sampling in the Fourier Space*

In the presence of long-range interactions such as the Coulomb interactions, the molecular dynamics simulation becomes computationally expensive for large N, especially when the periodic box is used to represent a system with large size. A lot of effort has already been devoted to computing such long-range interactions efficiently. Some popular methods include lattice summation methods such as the particle-particle mesh Ewald (PPPM) [76, 23], and multipole type methods such as treecode [7, 25] and fast multipole methods (FMM) [42, 101]. These methods can reduce the complexity per time step from $O(N^2)$ to $\mathcal{O}(N \log N)$ or even $\mathcal{O}(N)$, and have gained big success in practice. However, some issues still remain to be resolved, e.g., the prefactor in the linear scaling can be large, or the implementation can be nontrivial, or the scalability for parallel computing is not high.

In this section, we give a brief introduction to the recently proposed Random Batch Ewald (RBE) method for molecular dynamics simulations of particle systems with long-range Coulomb interactions in a periodic box, which achieves an $\mathcal{O}(N)$ complexity [59] with a high parallel efficiency. A natural splitting that breaks the Coulomb potential into the short range singular part and long-range smooth part as in Sect. 4.1 is the Ewald splitting. Due to the periodic setting, direct application of RBM to the long-range smooth part needs to group the particles chosen (with net charge zero) together with their periodic images. This works but the variance is not very small. The interesting observation in [59] is that the Ewald sum for the long-range part can be written in the Fourier space with a discrete Gaussian weight. Then, an important sampling mini-batch strategy can be used to reduce the variance significantly. In one sentence, the RBE method is based on the Ewald splitting for the Coulomb kernel with a random "mini-batch" type technique applied in the Fourier series for the long-range part.

Compared with PPPM where the Fast Fourier Transform is used to speed up the computation in the Fourier space, the RBE method uses random batch type technique to speed up the computation.

Consider N physical or numerical particles inside the periodic box with side length L, assumed to have net charge q_i $(1 \le i \le N)$ with the electroneutrality condition

$$\sum_{i=1}^{N} q_i = 0. \tag{4.4}$$

The forces are computed using $\boldsymbol{F}_i = -\nabla_{\boldsymbol{r}_i} U$, where U is the potential energy of the system. Since the Coulomb potential is of long range, with the periodic boundary condition, one must consider the images so that

$$U = \frac{1}{2} \sum_{\boldsymbol{n}}{}' \sum_{i,j=1}^{N} q_i q_j \frac{1}{|\boldsymbol{r}_{ij} + \boldsymbol{n} L|}, \tag{4.5}$$

where $\boldsymbol{n} \in \mathbb{Z}^3$ ranges over the three-dimensional integer vectors and $\sum'$ is defined such that $\boldsymbol{n} = 0$ is not included when $i = j$.

Due to the long-range nature of the Coulomb potential, the series (4.5) converges conditionally, thus a naive truncation would require a very large r to maintain the desired numerical accuracy. The classical Ewald summation separates the series into long-range smooth parts and short range singular parts:

$$\frac{1}{r} = \frac{\operatorname{erf}(\sqrt{\alpha} r)}{r} + \frac{\operatorname{erfc}(\sqrt{\alpha} r)}{r}, \tag{4.6}$$

where $\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \exp(-u^2)du$ is the error function and $\mathrm{erfc} = 1 - \mathrm{erf}$. Correspondingly,

$$U_1 = \frac{1}{2} \sum_{\boldsymbol{n}}{}' \sum_{i,j} q_i q_j \frac{\mathrm{erf}(\sqrt{\alpha}|\boldsymbol{r}_{ij} + \boldsymbol{n}L|)}{|\boldsymbol{r}_{ij} + \boldsymbol{n}L|}, \tag{4.7}$$

$$U_2 = \frac{1}{2} \sum_{\boldsymbol{n}}{}' \sum_{i,j} q_i q_j \frac{\mathrm{erfc}(\sqrt{\alpha}|\boldsymbol{r}_{ij} + \boldsymbol{n}L|)}{|\boldsymbol{r}_{ij} + \boldsymbol{n}L|}. \tag{4.8}$$

The computation of force can be done directly using

$$\boldsymbol{F}_i = -\nabla_{\boldsymbol{r}_i} U = -\nabla_{\boldsymbol{r}_i} U_1 - \nabla_{\boldsymbol{r}_i} U_2 =: \boldsymbol{F}_{i,1} + \boldsymbol{F}_{i,2}.$$

The second part $\boldsymbol{F}_{i2}$ corresponds to the short range forces whose computational cost is relatively low, since, for each particle, one just needs to add a finite number of particles in its close neighborhood. We now focus on the first part.

The slow decay of U_1 in r, corresponding to the long-range, can be dealt with in the Fourier space thanks to its smoothness (see [34, Chap. 12]):

$$U_1 = \frac{2\pi}{V} \sum_{\boldsymbol{k}\neq 0} \frac{1}{k^2} |\rho(\boldsymbol{k})|^2 e^{-k^2/4\alpha} - \sqrt{\frac{\alpha}{\pi}} \sum_{i=1}^{N} q_i^2, \tag{4.9}$$

where $k = |\boldsymbol{k}|$ and $\rho(\boldsymbol{k})$ is given by $\rho(\boldsymbol{k}) := \sum_{i=1}^N q_i e^{i\boldsymbol{k}\cdot\boldsymbol{r}_i}$. The divergent $\boldsymbol{k} = 0$ term is usually neglected in simulations to represent that the periodic system is embedded in a conducting medium which is essential for simulating ionic systems. Then

$$\boldsymbol{F}_{i,1} = -\sum_{\boldsymbol{k}\neq 0} \frac{4\pi q_i \boldsymbol{k}}{V k^2} e^{-k^2/(4\alpha)} \mathrm{Im}(e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_i} \rho(\boldsymbol{k})), \tag{4.10}$$

where we recall $\boldsymbol{r}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i$, pointing toward particle j from particle i. Note that the force $\boldsymbol{F}_{i,1}$ is bounded for small $\boldsymbol{k}$. In fact, $k \geq 2\pi/L$, so $Vk \geq 2\pi L^2$. Let us consider the factor $e^{-k^2/(4\alpha)}$, and denote the sum of such factors by

$$S := \sum_{\boldsymbol{k}\neq 0} e^{-k^2/(4\alpha)} = H^3 - 1, \tag{4.11}$$

where

$$H := \sum_{m\in\mathbb{Z}} e^{-\pi^2 m^2/(\alpha L^2)} = \sqrt{\frac{\alpha L^2}{\pi}} \sum_{m=-\infty}^{\infty} e^{-\alpha m^2 L^2} \approx \sqrt{\frac{\alpha L^2}{\pi}} (1 + 2e^{-\alpha L^2}), \tag{4.12}$$

Algorithm 6 (Random batch Ewald)

```
Choose $\alpha$, $r_c$, and $k_c$ (the cutoffs in real and Fourier spaces respectively), $\Delta t$, and batch size $p$. Initialize the positions and velocities of charges $\boldsymbol{r}_i^0$, $\boldsymbol{v}_i^0$ for $1 \le i \le N$.
Sample sufficient number of $\boldsymbol{k} \sim e^{-k^2/(4\alpha)}$, $\boldsymbol{k} \neq 0$ by the MH procedure to form a set $\mathcal{K}$.
for $n$ in $1 : N$ do
    Integrate Newton's equations (4.1) for time $\Delta t$ with appropriate integration scheme and some appropriate thermostat. The Fourier parts of the Coulomb forces are computed using RBE force (4.14) with the $p$ frequencies chosen from $\mathcal{K}$ in order.
end for
```

since often $\alpha L^2 \gg 1$. Hence, S is the sum for all three-dimensional vectors $\boldsymbol{k}$ except 0. Then, one can regard the sum as an expectation over the probability distribution

$$\mathscr{P}_{\boldsymbol{k}} := S^{-1} e^{-k^2/(4\alpha)}, \tag{4.13}$$

which, with $\boldsymbol{k} \neq 0$, is a discrete Gaussian distribution and can be sampled efficiently. For example, one can use the Metropolis-Hastings (MH) algorithm (see [48] for details) by choosing proposal samples from the continuous Gaussian $\mathcal{N}(0, \alpha L^2/(2\pi^2))$, the normal distribution with mean zero and variance $\alpha L^2/(2\pi^2)$. It should be emphasized that this sampling can be done *offline*, before the iteration begins. Once the time evolution starts one just needs to randomly draw a few (p) samples for each time step from this pre-sampled Gaussian sequence.

With this observation, the MD simulations can then be done via the random mini-batch approach with this importance sampling strategy. Specifically, one approximates the force $\boldsymbol{F}_{i,1}$ in (4.10) by the following random variable:

$$\boldsymbol{F}_{i,1} \approx \boldsymbol{F}_{i,1}^* := -\sum_{\ell=1}^{p} \frac{S}{p} \frac{4\pi \boldsymbol{k}_\ell q_i}{V k_\ell^2} \operatorname{Im}(e^{-i\boldsymbol{k}_\ell \cdot \boldsymbol{r}_i} \rho(\boldsymbol{k}_\ell)). \tag{4.14}$$

The corresponding algorithm is shown in Algorithm 6.

Similar to the strategy in the PPPM, one may choose α such that the time cost in real space is cheap and then speed up the computation in the Fourier space. Compared with PPPM, the only difference is that PPPM uses FFT while RBE uses random mini-batch to speed up the computation in the Fourier space. Hence, we make the same choice

$$\sqrt{\alpha} \sim \frac{N^{1/3}}{L} = \rho_r^{1/3},$$

which is inverse of the average distance between two numerical particles. The complexity for the real space part is $\mathcal{O}(N)$. By choosing *the same batch* of frequencies for all forces (4.14) (i.e., using the same $\boldsymbol{k}_\ell$, $1 \le \ell \le p$ for all $\boldsymbol{F}_{i,1}^*$, $1 \le i \le N$) in the same time step, the complexity per iteration for the frequency part is reduced to $\mathcal{O}(pN)$. This implies that the RBE method has linear complexity per time step if one chooses $p = \mathcal{O}(1)$.

To illustrate the performance of the RBE method, consider an electrolyte with monovalent binary ions (first example in [59]). In the reduced units ([34, section 3.2]), the dielectric constant is taken as $\varepsilon = 1/4\pi$ so that the potential of a charge is $\phi(r) = q/r$ and the temperature is $T = \beta^{-1} = 1$. Under the Debye-Hückel (DH) theory (linearized Poisson-Boltzmann equation), the charge potential outside one ion is given by

$$-\varepsilon\Delta\phi = \begin{cases} 0 & r < a \\ q\rho_{\infty,+}e^{-\beta q\phi} - q\rho_{\infty,-}e^{\beta q\phi} \approx \beta q^2 \rho_r \phi, & r > a \end{cases}$$

where $\rho_{\infty,+} = \rho_{\infty,-} = N/(2V)$ are the densities of the positive and negative ions at infinity, both being $\rho_r/2$. The parameter a is the effective diameter of the ions, which is related to the setting of the Lennard-Jones potential (in real simulations, besides the Coulomb interactions computed using the RBE method, the Lennard-Jones potential is also considered). In the simulations, $a = 0.2$ and the setting of Lennard-Jones potential can be found in [59]. This approximation gives the net charge density $\rho = -\varepsilon\Delta\phi$ for $r \gg a$,

$$\ln(r\rho(r)) \approx -1.941r - 1.144.$$

The results in the left panel of Fig. 4 were obtained by $N = 300$ (i.e., 150 cation and anion particles respectively) in a periodic box with side length $L = 10$. The thermostat was Andersen's thermostat with collision frequency $\nu = 3$. These parameters are chosen such that they give comparable results. Clearly, all the three methods give correct results, agreeing with the curve predicted by the DH theory. Regarding the efficiency, the right figure shows the time consumed for different

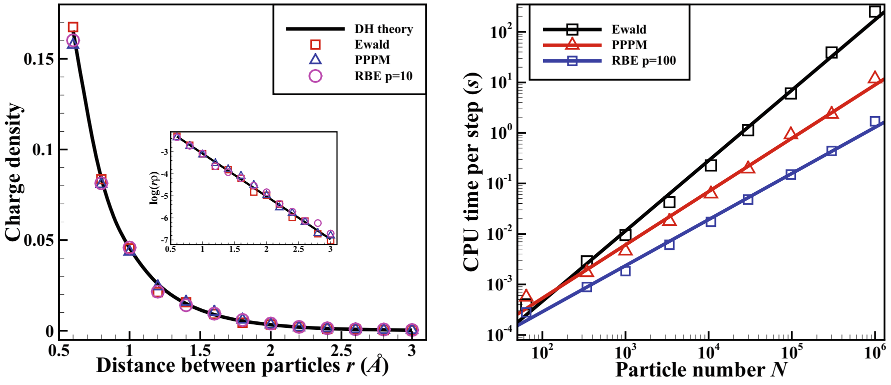


Fig. 4 Comparison of the Ewald sum, the PPPM and the RBE methods

Table 1 Relative error of potential energy for the RBE method against PPPM method with different densities and batch sizes

	$p = 10$	$p = 20$	$p = 50$	$p = 100$
$\rho_r = 0.1$	0.15%	0.13%	0.13%	0.08%
$\rho_r = 0.3$	0.10%	0.08%	0.04%	0.09%
$\rho_r = 1$	0.66%	0.18%	0.11%	0.04%
$\rho_r = 4$	7.83%	2.38%	0.71%	0.31%

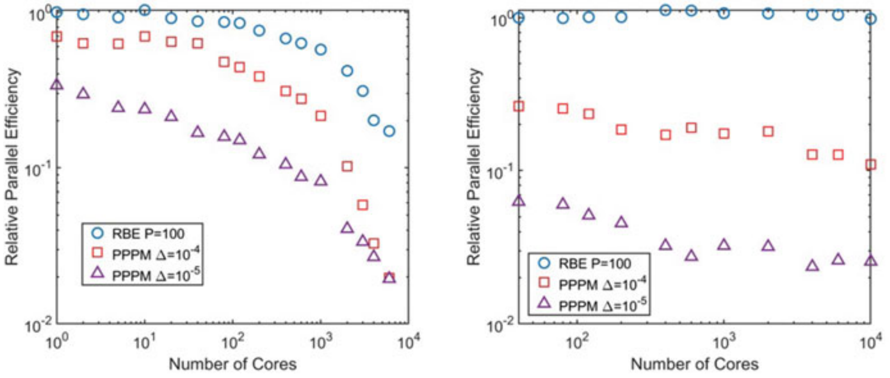


Fig. 5 The parallel efficiency of the PPPM and the RBE methods for all-atom simulation of pure water system (Left) 3×10^5 atoms; (Right) 3×10^7 atoms

particle numbers inside the box with the same side length $L = 10$. Both the PPPM and RBE methods scale linearly with the particle numbers. However, even for batch size $p = 100$, the RBE method consumes much less time. The relative accuracies of the potential obtained by RBE against the PPPM are listed in Table 1, for different densities $\rho_r = N/L^3$. Clearly, the RBE method has the same level of accuracy compared with the PPPM method for the densities considered.

Next, in Fig. 5, the parallel efficiency of the PPPM and RBE methods from [71] for the all-atom simulation of pure water systems is shown. As can be seen, due to the reduction of communications for the particles, the RBE method gains better parallel efficiency. This parallel efficiency is more obvious when the number of particles is larger. In [71], the simulation results of pure water system also indicate that the RBE type methods can not only sample from the equilibrium distribution, but also compute accurately the dynamical properties of the pure water systems.

5 Statistical Sampling

Sampling from a complicated or even unknown probability distribution is crucial in many applications, including numerical integration for statistics of many-body systems [34, 29], parameter estimation for Bayesian inference [93, 12], etc.. The methods that rely on random numbers for sampling and numerical simulations are generally called the Monte Carlo (MC) methods [61, 29]. The law of large numbers [27] validates the usage of empirical measures for approximation of the complicated or unknown probability measure. By the central limit theorem [27], the error of the MC methods scales like $\mathcal{O}(N^{-1/2})$ which is independent of the dimension d, hence the MC methods overcome the curse of dimensionality. The Markov Chain Monte Carlo (MCMC) methods [39, 35] are among the most popular MC methods. By constructing Markov chains that have the desired distributions to be the invariant measures, one can obtain samples from the desired distributions by recording the states of the Markov chains. A typical MCMC algorithm is the Metropolis-Hastings algorithm [81, 48].

Unlike the MCMC, the Stein variational Gradient method (proposed by Liu and Wang in [74]) belongs to the class of *particle based* variational inference sampling methods (see also [88, 20]). These methods update particles by solving optimization problems, and each iteration is expected to make progress toward the desired distribution. As a non-parametric variational inference method, SVGD gives a deterministic way to generate points that approximate the desired probability distribution by solving an ODE particle system, which displays different features from the Monte Carlo methods.

We describe in this section two sampling methods that use RBM to improve the efficiency. The first method, Random Batch Monte Carlo, is a fast MCMC that costs only $\mathcal{O}(1)$ per iteration to sample from the Gibbs measures corresponding to many-body particle systems with singular interacting kernels. The second method, RBM-SVGD, is an interesting application of RBM to the Stein variational gradient descent ODE system, which is an interacting particle system.

5.1 Random Batch Monte Carlo for Many-Body Systems

Suppose that one wants to sample from the N-particle Gibbs distribution

$$\pi(\underline{x}) \propto \exp\left[-\beta H(\underline{x})\right], \tag{5.1}$$

with $\underline{x} = (x_1, \cdots, x_N) \in \mathbb{R}^{Nd}$ ($x_i \in \mathbb{R}^d$, and $d \geq 1, d \in \mathbb{N}$), β being a positive constant, the N-body energy

$$H(\underline{x}) := \sum_{i=1}^{N} w_i V(x_i) + \sum_{i,j:i<j} w_i w_j \phi(x_i - x_j), \tag{5.2}$$

and V being the external potential assumed to be smooth. Here, w_i's are the weights. In the molecular regime, w_i's are often taken to be 1, while in the mean-field regime [92, 37, 66], one may have $w \sim N^{-1}$.

In [70], Li et. al. proposed the Random Batch Monte Carlo method, which costs $\mathcal{O}(1)$ per time step for sampling from equilibrium distributions (Gibbs measures) corresponding to particle systems with singular interacting kernels. Similarly to [77, 49] and the MD methods above, the interacting potential is decomposed into two parts

$$\phi(x) = \phi_1(x) + \phi_2(x), \tag{5.3}$$

where we suppose that ϕ_1 has long range but is smooth and bounded, while ϕ_2 is singular and of short range. The algorithm is based on the following splitting Monte Carlo, which is a special case of the Metropolis-Hastings algorithm:

Suppose there are N particles located at x_j for $j = 1, \cdots, N$. Let us consider the following method for a Markovian jump.

Step 1 —Randomly choose a particle i.

Step 2 —Move the particle using ϕ_1 with overdamped Langevin equation:

$$\begin{aligned} d\boldsymbol{r}_i &= -\left(\frac{\nabla V(\boldsymbol{r}_i)}{w(N-1)} + \frac{1}{N-1} \sum_{j:j\neq i} \nabla\phi_1(\boldsymbol{r}_i - \boldsymbol{r}_j) \right) dt + \sqrt{\frac{2}{(N-1)w^2\beta}}\, d\boldsymbol{W}_i, \\ \boldsymbol{r}_i(0) &= x_i, \end{aligned} \tag{5.4}$$

where x_j's are fixed. Evolve this SDE with some time $t > 0$ and obtain $\boldsymbol{r}_i(t) \to x_i^*$ as a candidate position of particle i for the new sample.

Step 3 —Use ϕ_2 to do the Metropolis rejection. Define

$$\mathrm{acc}(x_i, x_i^*) = \min\left\{ 1, \exp\Big[-\beta \sum_{j:j\neq i} w^2(\phi_2(x_i^* - x_j) - \phi_2(x_i - x_j)) \Big] \right\}. \tag{5.5}$$

With probability $\mathrm{acc}(x_i, x_i^*)$, accept x_i^* and set

$$x_i \leftarrow x_i^*. \tag{5.6}$$

Otherwise, x_i is unchanged. Then, a new sample $\{x_1, \cdots, x_N\}$ is obtained for the Markov chain.

Note that the overdamped Langevin equation satisfies the detailed balance condition so the above algorithm is a special case of the Metropolis-Hastings algorithm, thus can correctly sample from the desired Gibbs distribution. Due to the short range of ϕ_2, *Step 3* can be done in $\mathcal{O}(1)$ operations using some standard data structures such as the cell list [34, Appendix F]. The idea is to use the random mini-batch approach to *Step 2*. Hence, one discretizes the SDE with the Euler-Maruyama scheme [63, 82]. The interaction force is approximated within the random mini-batch idea. This gives the following algorithm.

Algorithm 7 (Random batch Monte Carlo algorithm)

```
Split $\phi := \phi_1 + \phi_2$ such that $\phi_1$ is smooth and with long range; $\phi_2$ is with short range. Generate
$N$ initial particles; choose $N_s$ (the total number of samples), $p > 1, m \geq 1$
for $n$ in $1 : N_s$ do
    Randomly pick an index $i \in \{1, \cdots, N\}$ with uniform probability
    $\boldsymbol{r}_i \leftarrow x_i$
    for $k = 1, \cdots, m$ do
        Choose $\boldsymbol{\xi}_k, z_k \sim \mathcal{N}(0, I_d)$, $\Delta t_k > 0$ and let,

        $$\boldsymbol{r}_i \leftarrow \boldsymbol{r}_i - \Delta t_k \left[ \frac{\nabla V(\boldsymbol{r}_i)}{w(N-1)} + \frac{1}{p-1} \sum_{j \in \boldsymbol{\xi}_k} \nabla \phi_1(\boldsymbol{r}_i - x_j) \right] + \sqrt{\frac{2\Delta t_k}{(N-1)w^2\beta}} z_k$$

    end for
    Let $x_i^* \leftarrow \boldsymbol{r}_i$. Compute the following using cell list or other data structures:

    $$\alpha = \min \left\{ 1, \exp \Big[ -\beta \sum_{j:j\neq i} w^2(\phi_2(x_i^* - x_j) - \phi_2(x_i - x_j)) \Big] \right\}$$

    Generate a random number $\zeta$ from uniform distribution on $[0, 1]$. If $\zeta \leq \alpha$, set

    $$x_i \leftarrow x_i^*$$

end for
```

It has been proved in [70] that the mini-batch approximation has an error control for the transition probability so that the method is correct with some systematic error. The computational cost is $\mathcal{O}(1)$ for each iteration and the efficiency could be higher since there is no rejection in *Step 2*.

We now present a numerical result from [70] to illustrate the efficiency of RBMC. Consider the Dyson Brownian motion [31]:

$$d\lambda_j(t) = -\lambda_j(t)\,dt + \frac{1}{N-1} \sum_{k:k\neq j} \frac{1}{\lambda_j - \lambda_k} dt + \frac{1}{\sqrt{N-1}} dW_j, \quad j = 1, \cdots, N, \tag{5.7}$$

where $\{\lambda_j\}$'s represent the eigenvalues of certain random matrices (compared with the original Dyson Brownian motion, $N-1$ instead of N is used in (5.7); there is little effect due to the replacement $N \to N-1$). In the limit $N \to \infty$, the distribution obeys the following nonlocal PDE

$$\partial_t \rho(x,t) + \partial_x(\rho(u - x)) = 0, \quad u(x,t) = \pi(H\rho)(x,t) = \text{p.v.} \int_{\mathbb{R}} \frac{\rho(y,t)}{x-y}\, dy, \tag{5.8}$$

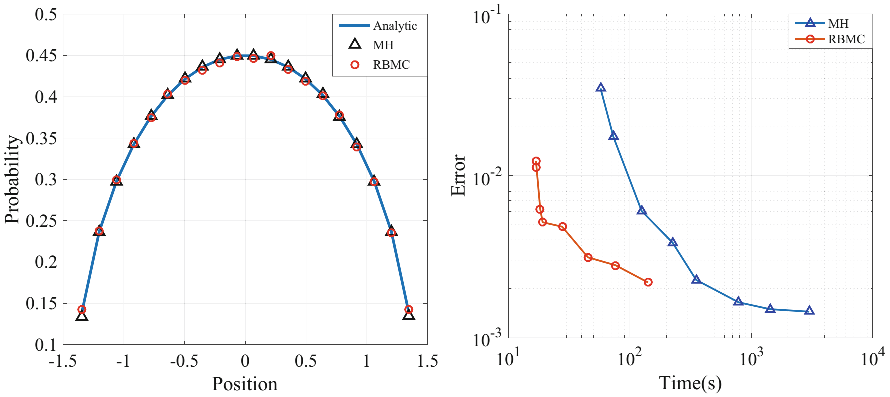


Fig. 6 (Left) Empirical densities with 1e7 sampling iterations (1e7N sample points). The blue curve is the analytical curve given by the semicircle law (5.9). (Right) error versus CPU time

where $H(\cdot)$ is the Hilbert transform on $\mathbb{R}$, $\pi = 3.14\cdots$ is the circumference ratio and p.v. represents the Cauchy principal value. From this PDE, one finds that the limiting equation (5.8) has an invariant measure, given by the semicircle law:

$$\rho(x) = \frac{1}{\pi}\sqrt{2-x^2}. \tag{5.9}$$

Figure 6 shows the sampling results of RBMC and MH methods for empirical measures with particles from the joint distribution

$$\pi(d\underline{x}) \propto \exp\left(-\Big(\frac{N-1}{2}\sum_i x_i^2 - \sum_{i<j}\ln|x_i - x_j|\Big)\right),$$

which is the invariant measure for the interacting particle system (5.7). The empirical measure is expected to be close to the semicircle law when N is large enough. In the simulations, the particle number was fixed as $N = 500$. In the RBMC, the splitting was done for $\ln r$ at $r = 0.01$, and the time step was chosen as $\Delta t = 10^{-4}$. The MH algorithm uses a certain Gaussian proposal for the random movement of a chosen particle. The left panel of Fig. 6 shows that both methods yield results that agree with the semicircle law reasonably well. The right panel plots the relative error with respect to the semicircle law versus CPU time. Clearly, the RBMC method only needs 10% of the time for the MH method to get the error tolerance considered.

5.2 RBM-SVGD: A Stochastic Version of Stein Variational Gradient Descent

Suppose that one is interested in some target probability distribution with density $\pi(x)$ $(x \in \mathbb{R}^d)$. In SVGD, one sets $V = -\log \pi$, chooses some symmetric positive definite kernel $\mathcal{K}(x, y)$, and solves the following ODE system for given initial points $\{\boldsymbol{r}_i(0)\}_{i=1}^N$ (see [74, 73]):

$$\dot{\boldsymbol{r}}_i = \frac{1}{N}\sum_{j=1}^{N} \nabla_y \mathcal{K}(\boldsymbol{r}_i, \boldsymbol{r}_j) - \frac{1}{N}\sum_{j=1}^{N} \mathcal{K}(\boldsymbol{r}_i, \boldsymbol{r}_j)\nabla V(\boldsymbol{r}_j), \quad i = 1, \cdots, N, \tag{5.10}$$

where N is the number of particles for the sampling purpose. The subindex "y" in ∇_y means that the gradient is taken with respect to the second variable in $\mathcal{K}(\cdot, \cdot)$; i.e., $\nabla_y \mathcal{K}(\boldsymbol{r}_i, \boldsymbol{r}_j) := \nabla_y \mathcal{K}(x, y)|_{(x,y)=(\boldsymbol{r}_i, \boldsymbol{r}_j)}$. When t is large enough, the empirical measures constructed using $\{\boldsymbol{r}_i(t)\}_{i=1}^N$ is expected to be close to π, i.e.,

$$\frac{1}{N}\sum_{i=1}^{N} \delta(x - \boldsymbol{r}_i(t)) \approx \pi(x)\, dx, \quad t \gg 1.$$

SVGD provides consistent estimation for generic distributions as Monte Carlo methods do, but it seems to be more efficient than some Monte Carlo methods in practice level for approximating the desired measure, when the number of particles is small [74, 24]. Interestingly, it reduces to the maximum a posterior (MAP) method when $N = 1$ [74].

The ODE system (5.10) clearly is an interacting particle system but now the interaction kernel is no longer translation invariant and is not symmetric. The kernel can even grow as $|\boldsymbol{r}_i - \boldsymbol{r}_j| \to \infty$. Clearly, for such systems, RBM is applicable. Applying the RBM to this special kernel and using any suitable ODE solvers, one gets a class of sampling algorithms, which is called RBM-SVGD in [67]. The discrete algorithm (with possible variant step size) is shown in Algorithm 8. Clearly, the complexity is $\mathcal{O}(pN)$ for each iteration.

Here, N_T is the number of iterations and $\{\eta_k\}$ is the sequence of time steps, which play the same role as learning rate in SGD [11, 13]. For some applications, one may simply set $\eta_k = \eta \ll 1$ to be a constant and get relatively good results. However, in many high dimensional problems, choosing η_k to be constant may yield divergent sequences [89]. One may decrease η_k to obtain convergent data sequences. For example, one may simply choose $\eta_k = 1/k$ as in SGD. Another frequently used strategy is the AdaGrad approach [26, 97].

We recall the gradient flow under the so-called Stein metric in the space of probability measures [73, 36]:

Algorithm 8 RBM-SVGD

```
for $k$ in $0 : N_T - 1$ do
    Divide $\{1, 2, \dots, pn\}$ into $n$ batches randomly.
    for each batch $\mathcal{C}_q$ do
        For all $i \in \mathcal{C}_q$,
        $$\boldsymbol{r}_i^{(k+1)} \leftarrow \boldsymbol{r}_i^{(k)} + \frac{1}{N}\Big(\nabla_y \mathcal{K}(\boldsymbol{r}_i^{(k)}, \boldsymbol{r}_i^{(k)}) - \mathcal{K}(\boldsymbol{r}_i^{(k)}, \boldsymbol{r}_i^{(k)}) \nabla V(\boldsymbol{r}_i^{(k)})\Big)\eta_k + \Phi_{k,i}\eta_k,$$
        where
        $$\Phi_{k,i} = \frac{N-1}{N(p-1)} \sum_{j \in \mathcal{C}_q, j \neq i} \Big(\nabla_y \mathcal{K}(\boldsymbol{r}_i^{(k)}, \boldsymbol{r}_j^{(k)}) - \mathcal{K}(\boldsymbol{r}_i^{(k)}, \boldsymbol{r}_j^{(k)}) \nabla V(\boldsymbol{r}_j^{(k)})\Big). \quad (5.11)$$
    end for
end for
```

$$\partial_t \rho = \nabla \cdot \left(\rho \mathcal{K} * (\rho \nabla \frac{\delta E}{\delta \rho}) \right), \tag{5.12}$$

where $\mathcal{K} * g = \int \mathcal{K}(x, y) g(y)\, dy$. Consider taking the energy functional as the Kullback–Leibler (KL) divergence between ρ and the target distribution π, where KL divergence is also known as the relative entropy defined by

$$\mathrm{KL}(\mu || \nu) = \mathbb{E}_{Y \sim \mu} \log \left(\frac{d\mu}{d\nu}(Y) \right). \tag{5.13}$$

Here $\frac{d\mu}{d\nu}$ is the well-known Radon–Nikodym derivative. Then, Eq. (5.12) becomes

$$\partial_t \rho = \nabla \cdot (\rho \mathcal{K} * (\rho \nabla V + \nabla \rho)). \tag{5.14}$$

It is easy to see that $\pi \propto \exp(-V)$ is invariant under this PDE. See [73, 75] for some relevant studies.

The above theory encounters difficulty for empirical measures because the KL divergence is simply infinity. One benefit of the of the "Stein metric" is that the gradient may be moved from $\nabla \rho$ onto the kernel $\mathcal{K}(x, y)$ so that the flow (5.12) becomes (5.10), which is then well-defined. In fact, if $\{\boldsymbol{r}_i\}$ solves the ODE system (5.10), then the corresponding empirical measure is a measure solution to (5.14) (see [75, Proposition 2.5]). Hence, one may reasonably expect that (5.10) will give approximation for the desired distribution π.

For numerical illustration, we take an example from [67]. Consider the logistic regression for binary classification on the Covertype dataset, with 581012 data points and 54 features [38]. The inference is applied on posterior $p(x|D)$ with the parameter $x = [w, \log \alpha]$ being of dimension 55. Here, D is 80% of the data and the remaining data were used for test. Figure 7 shows the performance of SVGD and

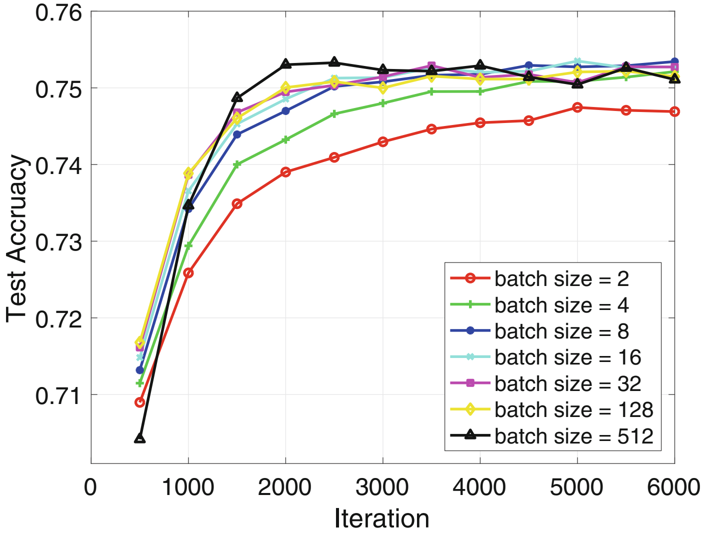


Fig. 7 Test accuracy on the Covertype dataset

Table 2 Average runtime of 6000 iterations

	RBM-SVGD						SVGD
p	2	4	8	16	32	128	512
Runtime(s)	8.59	11.24	16.28	26.15	21.66	19.42	47.01
Speedup	5.5x	4.2x	2.9x	1.8x	2.2x	2.4x	

RBM-SVGD with $N = 512$ particles and kernel $\mathcal{K}(x, y) = k(x - y)$ for a Gaussian kernel $k(\cdot)$. Clearly, RBM-SVGD gives comparable results with SVGD, both results being as good as some traditional methods.

Table 2 shows the CPU time and speedup of RBM-SVGD. Clearly, for comparable results, RBM-SVGD is more efficient.

6 Agent-Based Models for Collective Dynamics

Collective behaviors of self-propelled particles (agents) are ubiquitous in nature, for example, synchronous flashing of fireflies and pacemaker cells, swarming of fish, flocking of birds and herding of sheep. We refer to [1, 19, 83, 94, 100] for survey articles and related literature.

While the RBM was introduced as an efficient algorithm for interacting particle systems, one can also view it as a (random) model of the underlying problem, which takes into account only a small number of interactions randomly at discrete time steps. Two natural questions arise with such models: (a) How accurate are these "new" random models compared to the original, full batch models? (b) Do these random models still capture the main features of the original model, such as the collective or long-time behavior, and under what conditions? Here we review some recent results that address these issues for two representative problems, the Cucker–Smale model for flocking and the consensus model.

6.1 The Cucker–Smale Model

We begin with the Cucker–Smale (CS) model [19]:

$$\begin{cases} \dfrac{d\boldsymbol{x}_i}{dt} = \boldsymbol{v}_i, \quad t > 0, \quad i = 1, \dots, N, \\ \dfrac{d\boldsymbol{v}_i}{dt} = \dfrac{\kappa}{N-1} \displaystyle\sum_{j:j\neq i} \psi(|\boldsymbol{x}_j - \boldsymbol{x}_i|)(\boldsymbol{v}_j - \boldsymbol{v}_i), \\ (\boldsymbol{x}_i(0), \boldsymbol{v}_i(0)) = (x_i^{in}, v_i^{in}), \end{cases} \tag{6.1}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{v}_i$ are the position and velocity of the i-th CS particle, respectively, κ is the non-negative coupling strength and ψ, the communication weight measuring mutual interactions between agents, is positive, bounded, and Lipschitz continuous and satisfies the monotonicity conditions:

$$\begin{aligned} &0 \le \psi(r) \le \psi_M, \quad \forall\, r \ge 0, \quad \|\psi\|_{\mathrm{Lip}} < \infty, \\ &(\psi(r_1) - \psi(r_2))(r_1 - r_2) \le 0, \quad r_1,\ r_2 \in \mathbb{R}_+. \end{aligned} \tag{6.2}$$

Here $\psi_M > 0$ is a constant. The emergent dynamics of (6.1), flocking, in which all particles will eventually stay in a bounded domain with the same velocity, has been extensively studied in literature [45, 47].

Consider the RBM approximation for (6.1):

$$\begin{cases} \dfrac{d\tilde{\boldsymbol{x}}_i}{dt} = \tilde{\boldsymbol{v}}_i, \quad t \in [t_{m-1}, t_m),\ m = 1, 2, \cdots, \\ \dfrac{d\tilde{\boldsymbol{v}}_i}{dt} = \dfrac{\kappa}{p-1} \displaystyle\sum_{j\in\mathcal{C}_i^{(k)}, j\neq i} \psi(|\tilde{\boldsymbol{x}}_j - \tilde{\boldsymbol{x}}_i|)(\tilde{\boldsymbol{v}}_j - \tilde{\boldsymbol{v}}_i), \\ (\tilde{\boldsymbol{x}}_i(0), \tilde{\boldsymbol{v}}_i(0)) = (x_i^{in}, v_i^{in}), \quad i = 1, \dots, N. \end{cases} \tag{6.3}$$

Assume that ψ is long-ranged:

$$1/\psi(r) = \mathcal{O}(r^{\beta}) \quad \text{as } r \to \infty \quad \text{for some } \beta \in [0, 1). \tag{6.4}$$

For example, one can take

$$\psi(r) = \frac{1}{(1+r^2)^{\beta/2}}, \quad \beta \in [0,1).$$

Then [46] establishes the following emergence of a global flocking: there exist positive constants $\tilde{x}_\infty$ and C such that

$$\begin{aligned}
&\sup_{0\le t<\infty} \mathbb{E}\Big(\frac{1}{N^2}\sum_{i,j=1}^N |\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{x}}_j|^2\Big) < \tilde{x}_\infty \quad \text{and} \\
&\mathbb{E}\Big(\frac{1}{N^2}\sum_{i,j=1}^N |\tilde{\boldsymbol{v}}_i - \tilde{\boldsymbol{v}}_j|^2\Big) \le C\exp\Big[-\frac{C(p-1)}{(N-1)(1+\Delta t)}t(1+t)^{-\beta}\Big],
\end{aligned} \tag{6.5}$$

where C depends only on ψ, β, κ and the initial data.

Furthermore, the following uniform-in-time error estimate was also proved: when ψ has a positive lower bound ψ_0,

$$\psi(r) \ge \psi_0 \quad \text{for } r \ge 0, \tag{6.6}$$

then

$$\begin{aligned}
\mathbb{E}\Big(\frac{1}{N}\sum_{i=1}^N |\tilde{\boldsymbol{v}}_i(t) - \boldsymbol{v}_i(t)|^2\Big) &\le C\Delta t\left(\frac{1}{p-1} - \frac{1}{N-1}\right) + C\Delta t^2 \\
&\quad + C(1+\Delta t)\exp(-\kappa\psi_0 t),
\end{aligned} \tag{6.7}$$

where the dependency of the constant C is the same as in (6.5).

Note that the positive lower bound assumption (6.6) corresponds to the case of $\beta = 0$ in the long-ranged communication (6.4). However, the third time-decaying term in the right-hand side of (6.7) is independent of p and N.

6.2 Consensus Models

Let $\boldsymbol{q}_i \in \mathbb{R}^d$, $1 \le i \le N$ be a collection of agents that seek for a consensus, governed by the Cauchy problem:

$$\begin{cases}
\dfrac{d\boldsymbol{q}_i}{dt} = \nu_i + \dfrac{\kappa}{N-1}\displaystyle\sum_{j\neq i} a_{ij}\Gamma(\boldsymbol{q}_j - \boldsymbol{q}_i), \quad t>0, \\
\boldsymbol{q}_i(0) = q_i^{in}, \quad i = 1,\cdots,N,
\end{cases} \tag{6.8}$$

where κ is a non-negative coupling strength and ν_i is the intrinsic velocity of the i-th agent. Here Γ is an interaction function satisfying the following properties: there exists $C_1 > 0$ such that

$$\Gamma \in \mathcal{C}^2(B_{C_1}(0)), \quad \Gamma(-q) = -\Gamma(q), \qquad \forall\, q \in \overline{B_{C_1}(0))}. \tag{6.9}$$

Here $B_r(x)$ is the open ball with radius r and center x. We assume, without loss of generality, that the total sum is zero:

$$\sum_{i=1}^{N} \nu_i = 0,$$

and the adjacency matrix $(a_{ij})_{i,j=1}^{N}$ represents the network structure for interactions between agents satisfying symmetry and non-negative conditions:

$$a_{ij} = a_{ji} \geq 0, \quad 1 \leq i, j \leq N.$$

Note that the first term on the R.H.S. of (6.8) induces the "*dispersion effect*" due to the heterogeneity of ν_i. The second term in the R.H.S. of (6.8), modeled by the convolution type consensus force, generates "*concentration effect*." The overall dynamics of (6.8) is determined by the competitions between dispersion and concentration.

Below we present the study on RBM to this problem in [64]. Consider the RBM approximation where the interaction term is approximated by the random mini-batch at each time step. Then the relative state $|\tilde{\boldsymbol{q}}_i - \tilde{\boldsymbol{q}}_j|$ for RBM approximation can be *unbounded* even if the original relative state $|\boldsymbol{q}_i - \boldsymbol{q}_j|$ is uniformly bounded. Thus to balance dispersion and interaction in the RBM, one also needs to apply the RBM in the dispersion part as well. A sufficient framework leading to the uniform boundedness of relative states is to introduce suitable decomposition of the dispersion term ν_i as a sum of N-dispersion terms $\bar{\nu}_{ij}$:

$$\bar{\nu}_{ij} = -\bar{\nu}_{ji}, \quad \nu_i = \frac{\kappa}{N-1} \sum_{j=1}^{N} \bar{\nu}_{ij}, \quad i, j = 1, \ldots, N. \tag{6.10}$$

Then, the original Cauchy problem (6.8) is equivalent to the following problem:

$$\begin{cases} \dfrac{d\boldsymbol{q}_i}{dt} = \dfrac{\kappa}{N-1} \displaystyle\sum_{j \neq i} \Big(\nu_{ij} + a_{ij}\Gamma(\boldsymbol{q}_j - \boldsymbol{q}_i) \Big), \quad t > 0, \\ \boldsymbol{q}_i(0) = q_i^{in}, \quad i = 1, \cdots, N, \end{cases} \tag{6.11}$$

and the RBM samples dispersions and interactions proportionally,

$$\begin{cases} \dfrac{d\tilde{\boldsymbol{q}}_i}{dt} = \dfrac{\kappa}{p-1} \displaystyle\sum_{j \in \mathcal{C}_i^{(k)}, j \neq i} \big(\bar{\nu}_{ij} + a_{ij}\Gamma(\tilde{\boldsymbol{q}}_j - \tilde{\boldsymbol{q}}_i) \big), \quad t \in (t_k, t_{k+1}), \\ \tilde{\boldsymbol{q}}_i(0) = q_i^{in}, \quad i = 1, \ldots, N, \ \ k = 0, 1, 2, \ldots. \end{cases} \tag{6.12}$$

We first state the main result for the *one-dimensional* case. Assume that the coupling function Γ is strongly dissipative in the sense that

$$(\Gamma(q_1) - \Gamma(q)) \cdot (q_1 - q) \approx |q_1 - q|^2, \quad \forall\, q, q_1 \in [-C_1, C_1],$$

and also the full system (6.8) has an equilibrium $\Phi = (\phi_1, \cdots, \phi_N) \in (-C_1, C_1)^N$ with initial data sufficiently close to Φ. The main result is the following uniform error estimate, under the condition that the underlying network topology is connected strongly enough:

$$\sup_{0\le t<\infty} \left[\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}|\tilde{\boldsymbol{q}}_i(t) - \boldsymbol{q}_i(t)|^2\right] \lesssim \left[\Delta t\left(\frac{1}{p-1} - \frac{1}{N-1}\right) + \Delta t^2\right].$$

For the multi-dimensional setting with $\boldsymbol{q}_i \in \mathbb{R}^d$, the same error analysis can be obtained under one more extra assumption, which guarantees that the states $Q := (\boldsymbol{q}_1, \cdots, \boldsymbol{q}_N)$ and $\tilde{Q} := (\tilde{\boldsymbol{q}}_1, \cdots, \tilde{\boldsymbol{q}}_N)$ are confined in the symmetric interval.

Now, we give two main results on the emergent dynamics of (6.12) proved in [43]. Introduce two functionals for $\tilde{Q} = (\tilde{\boldsymbol{q}}_1, \cdots, \tilde{\boldsymbol{q}}_N)$:

$$\mathcal{M}_2(\tilde{Q}) := \frac{1}{N}\sum_{j=1}^{N} |\tilde{\boldsymbol{q}}_j|^2, \quad \mathcal{D}(\tilde{Q}) := \max_{1\le i,j\le N} |\tilde{\boldsymbol{q}}_i - \tilde{\boldsymbol{q}}_j|.$$

The first main result is concerned with the exponential decay of the second moment of $\hat{q}_i^R$: there exists a positive constant $\Lambda_1 = \Lambda_1(N, P, \tau, \kappa, L_1)$ satisfying

$$\mathbb{E}\Big(\mathcal{M}_2(\tilde{Q}(t))\Big) \le e^{-\Lambda_1 t}\mathbb{E}\Big(\mathcal{M}_2(\tilde{Q}(0))\Big), \quad t \ge 0.$$

The second main result deals with almost sure (a.s.) convergence of $\tilde{Q}$: there exists a positive constant $\Lambda_2 = \Lambda_2(N, P, \tau, \kappa, L_1, L_2)$ such that

$$\mathcal{D}(\tilde{Q}(t)) \le \mathcal{D}(\tilde{Q}(0))Ce^{-\Lambda_2 t}, \quad t \ge 0.$$

We remark that although the exponential decay rates in above results depend on N, numerical results in [43] show that the decay rates are in not sharp, and they are independent of N.

7 Quantum Dynamics

In this section, we have a review of the applications of RBM to interacting particles in the quantum regime. In particular, we first present and comment on the convergence results of RBM applied to the N-body Schrödinger equation in [41], and then have a review of the application of RBM to quantum Monte Carlo (QMC) methods in [57].

7.1 A Theoretical Result on the N-Body Schrödinger Equation

The first principle computation is based on solving for complex-valued wave function $\Psi_N \equiv \Psi_N(t, x_1, \ldots, x_N) \in \mathbb{C}$ of the N-body Schrödinger equation

$$i\hbar\partial_t \Psi_N(t, x_1, \ldots, x_N) = \mathcal{H}_N \Psi_N(t, x_1, \ldots, x_N), \quad \Psi_N\big|_{t=0} = \Psi_N^{in}, \tag{7.1}$$

where $t \geq 0$ is the time while $x_m \in \mathbb{R}^d (1 \leq m \leq N)$ is the position of the mth particle, $\mathcal{H}_N$ is the quantum Hamiltonian for N identical particles with unit mass:

$$\mathcal{H}_N := \sum_{m=1}^{N} -\tfrac{1}{2}\hbar^2 \Delta_{x_m} + \tfrac{1}{N-1} \sum_{1 \leq \ell < n \leq N} V(x_\ell - x_n), \tag{7.2}$$

while $\hbar$ is the reduced Planck constant. The N particles in this system interact via a binary (real-valued) potential V assumed to be even, bounded, and sufficiently regular (at least of class $C^{1,1}$ on $\mathbb{R}^d$). The coupling constant $\frac{1}{N-1}$ is chosen in order to balance the summations in the kinetic energy (involving N terms) and in the potential energy (involving $\frac{1}{2}N(N-1)$ terms).

When solving (7.1), the computation is exceedingly expensive due to the smallness of $\hbar$, which requires small time steps Δt and small mesh sizes of order $\hbar$ for the convergence of the numerical scheme, due to the oscillation in the wave function Ψ_N with frequency of order $1/\hbar$ (see [6, 58]). On top of this, any numerical scheme for (7.1) requires computing, at each time step, the sum of the interaction potential for each particle pair in the N-particle system, which needs $\mathcal{O}(N^2)$ operations. The RBM described below reduces the computational cost to $\mathcal{O}(N)$ per time step.

Below we follow the presentation of [41]. Assume for simplicity that $N \geq 2$ is an even integer. Let $\sigma_1, \sigma_2, \ldots, \sigma_j, \ldots$ be a mutually independent and uniformly distributed random sequence of permutations. Each permutation $\sigma \in \mathfrak{S}_N$ defines a partition of $\{1, \ldots, N\}$ into $N/2$ batches of two indices:

$$\{1, \ldots, N\} = \bigcup_{k=1}^{N/2} \{\sigma(2k-1), \sigma(2k)\}.$$

Set

$$\mathbf{T}_t(\ell, n) := \begin{cases} 1 & \text{if } \{\ell, n\} = \left\{\sigma_{[\frac{t}{\Delta t}]+1}(2k-1), \sigma_{[\frac{t}{\Delta t}]+1}(2k)\right\} \text{ for some } k \le \frac{N}{2}\,, \\ 0 & \text{otherwise,} \end{cases} \tag{7.3}$$

and consider the time-dependent random batch Hamiltonian

$$\widetilde{\mathcal{H}}_N(t) := \sum_{m=1}^{N} -\tfrac{1}{2}\hbar^2 \Delta_{x_m} + \sum_{1\le \ell < n \le N} \mathbf{T}_t(\ell, n) V(x_\ell - x_n)\,. \tag{7.4}$$

The RBM then solves the random batch Schrödinger equation

$$i\hbar\partial_t \widetilde{\Psi}_N(t, x_1, \dots, x_N) = \widetilde{\mathcal{H}}_N(t)\widetilde{\Psi}_N(t, x_1, \dots, x_N)\,, \quad \widetilde{\Psi}_N\big|_{t=0} = \widetilde{\Psi}_N^{in}\,. \tag{7.5}$$

Clearly, for each time step the cost of computing the interaction potential is reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$.

As we have seen, RBM is known to converge in the case of classical dynamics. It is therefore natural to seek an error estimate for the quantum RBM. The major difficulty here is to obtain an error estimate that is *independent* of $\hbar$ and N.

7.1.1 Mathematical Setting and Main Result

It will be more convenient to carry out the analysis on the corresponding von Neumann equations

$$i\hbar\partial_t R_N(t) = \mathcal{H}_N R_N(t) - R_N(t)\mathcal{H}_N =: [\mathcal{H}_N, R_N(t)]\,, \quad R_N(0) = R_N^{in}\,. \tag{7.6}$$

Here we denote $\mathfrak{H} := L^2(\mathbb{R}^d; \mathbb{C})$ and $\mathfrak{H}_N = \mathfrak{H}^{\otimes N} \simeq L^2((\mathbb{R}^d)^N; \mathbb{C})$ for each $N \ge 2$. The algebra of bounded operators on $\mathfrak{H}$ is denoted by $\mathcal{L}(\mathfrak{H})$, while $\mathcal{L}^1(\mathfrak{H}) \subset \mathcal{L}(\mathfrak{H})$ and $\mathcal{L}^2(\mathfrak{H})$ are respectively the two-sided ideals of trace-class and Hilbert-Schmidt operators on $\mathfrak{H}$. The operator norm of $A \in \mathcal{L}(\mathfrak{H})$ is denoted $\|A\|$. A density operator on $\mathfrak{H}$ is a trace-class operator R on $\mathfrak{H}$ such that

$$R = R^* \ge 0 \quad \text{and} \quad \operatorname{trace}_{\mathfrak{H}}(R) = 1\,.$$

The set of density operators on a separable Hilbert space H is henceforth denoted $\mathcal{D}(H)$.

The random batch von Neumann equation is

$$i\hbar\partial_t \widetilde{R}_N(t) = [\widetilde{\mathcal{H}}(t), \widetilde{R}_N(t)]\,, \qquad \widetilde{R}_N(0) = R_N^{in}\,. \tag{7.7}$$

In order to find an error estimate for the RBM that is independent of the particle number N, one first needs to define in terms of $R_N(t)$ and $\widetilde{R}_N(t)$ *quantities of interest* to be compared that are *independent of* N. A common practice when considering large systems of identical particles is to study the reduced density operators, which unfortunately does not lead to N-independent error estimates [41]. Assume that R_N^{in} has an integral kernel $r^{in} \equiv r^{in}(x_1, \dots, x_N; y_1, \dots, y_N)$ satisfying the symmetry

$$r^{in}(x_1, \dots, x_N; y_1, \dots, y_N) = r^{in}(x_{\sigma(1)}, \dots, x_{\sigma(N)}; y_{\sigma(1)}, \dots, y_{\sigma(N)}) \tag{7.8}$$

for each permutation $\sigma \in \mathfrak{S}_N$. Then, for each $t \geq 0$, the N-body density operator $R_N(t)$ solution of (7.6) satisfies the same symmetry, i.e., its integral kernel of the form $r(t; x_1, \dots, x_N; y_1, \dots, y_N)$ also satisfies

$$r(t; x_1, \dots, x_N; y_1, \dots, y_N) = r(t; x_{\sigma(1)}, \dots, x_{\sigma(N)}; y_{\sigma(1)}, \dots, y_{\sigma(N)}) \tag{7.9}$$

for each permutation $\sigma \in \mathfrak{S}_N$. The 1-particle reduced density operator of $R_N(t) \in \mathcal{D}(\mathfrak{H}_N)$ is $R_{N,\mathbf{1}}(t) \in \mathcal{D}(\mathfrak{H})$ defined by the integral kernel

$$r_{\mathbf{1}}(t, x, y) := \int_{(\mathbb{R}^d)^{N-1}} r(t; x, z_2, \dots, z_N; y, z_2, \dots, z_N) dz_2 \dots dz_N \,. \tag{7.10}$$

Even if R_N^{in} satisfies the symmetry (7.8), in general $\widetilde{R}_N(t)$ does not satisfy the symmetry analogous to (7.9) for $t > 0$ (with r replaced with $\widetilde{r}$, an integral kernel for $\tilde{R}_N(t)$) because the random batch potential

$$\sum_{1 \leq \ell < n \leq N} \mathbf{T}_t(\ell, n) V(x_\ell - x_n)$$

is not invariant under permutations of the particle labels. For that reason, the 1-particle reduced density operator of $\widetilde{R}_N(t)$ one needs is $\widetilde{R}_{N,\mathbf{1}}(t) \in \mathcal{D}(\mathfrak{H})$ defined for all $t > 0$ by the integral kernel

$$\widetilde{r}_{\mathbf{1}}(t, x, y) := \frac{1}{N} \sum_{j=1}^{N} \int_{(\mathbb{R}^d)^{N-1}} \widetilde{r}(t; Z_{j,N}[x], Z_{j,N}[y]) d\hat{Z}_{j,N} \,, \tag{7.11}$$

with the notation

$$Z_{j,N}[x] := z_1, \dots, z_{j-1}, x, z_{j+1} \dots, z_N \,, \quad d\hat{Z}_{j,N} = dz_1 \dots dz_{j-1} dz_{j+1} \dots dz_N \,.$$

(Obviously (7.11) holds with $r_{\mathbf{1}}$ and r in the place of $\widetilde{r}_{\mathbf{1}}$ and $\widetilde{r}$ respectively because of the symmetry (7.9).)

We also need to introduce the Wigner functions of the density operators $R_N(t)$ and $\widetilde{R}_N(t)$. Let $s \equiv s(x, y) \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$ be an integral kernel of operator

$S \in \mathcal{L}^2(\mathcal{H})$. Then the Wigner function of S is defined by the formula

$$W_\hbar[S](x,\cdot) := \tfrac{1}{(2\pi)^d}\mathcal{F}\big(y \mapsto s(x+\tfrac{1}{2}\hbar y, x-\tfrac{1}{2}\hbar y)\big) \quad \text{for a.e. } x \in \mathbb{R}^d\,, \tag{7.12}$$

where $\mathcal{F}$ is the Fourier transform on $L^2(\mathbb{R}^d)$.

For each integer $M \geq 1$, we also introduce the dual norm

$$\begin{aligned}&|||f|||_{-M}\\ &:= \sup\left\{\left|\iint_{\mathbb{R}^d\times\mathbb{R}^d} f(x,\xi)\overline{a(x,\xi)}dxd\xi\right| \;\middle|\; \begin{array}{c} a \in C_c(\mathbb{R}^d\times\mathbb{R}^d)\,, \quad \text{and} \\ \max\limits_{\substack{|\alpha|,||\leq M\\ |\alpha|+||>0}} \|\partial_x^\alpha\partial_\xi^{\boldsymbol{a}}\|_{L^\infty(\mathbb{R}^d\times\mathbb{R}^d)} \leq 1\end{array}\right\}.\end{aligned} \tag{7.13}$$

The main results in [41] are the following theorem.

Theorem 7.1 *Assume that $N \geq 2$ and that $V \in C(\mathbb{R}^d)$ is a real-valued function such that*

$$V(z) = V(-z)\,\textit{for all}\; z \in \mathbb{R}^d\,, \qquad \lim_{|z|\to+\infty} V(z) = 0\,,$$
$$\textit{and} \int_{\mathbb{R}^d}(1+|\omega|^2)|\mathcal{F}(V)(\omega)|d\omega < \infty\,.$$

Let $R_{N,\mathbf{1}}(t)$ and $\tilde{R}_{N,\mathbf{1}}(t)$ be the single-particle reduced density operators defined in terms of $R_N(t)$ and $R_N^R(t)$ respectively by (7.10). Then there exists a constant $\gamma_d > 0$ depending only on the dimension d of the configuration space such that, for each $t > 0$, one has

$$\begin{aligned}&|||W_\hbar[\mathbb{E}\tilde{R}_{N,\mathbf{1}}(t)] - W_\hbar[R_{N,\mathbf{1}}(t)]|||_{-[d/2]-3}\\ &\quad\leq 2\gamma_d\Delta t e^{6t\max(1,\sqrt{d}L(V))}\Lambda(V)(2+3t\Lambda(V)\max(1,\Delta t)+4\sqrt{d}L(V)t\Delta t)\,,\end{aligned} \tag{7.14}$$

where

$$L(V) := \tfrac{1}{(2\pi)^d}\int_{\mathbb{R}^d}|\omega|^2|\hat{V}(\omega)|d\omega\,, \qquad \Lambda(V) := \tfrac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\sum_{\mu=1}^{d}|\omega^\mu||\hat{V}(\omega)|d\omega\,,$$

with ω^ν the ν-th component of ω.

This error estimate gives an error *independent* of $\hbar$ and N. It was also pointed out in [41] that the error bound obtained in above theorem is small as $\Delta t \to 0$, *even for moderate values of N* for which the factor $\frac{1}{N-1}$ is insignificant. Therefore the result applies to N-body quantum Hamiltonians *without the* $\frac{1}{N-1}$ *normalization of the interaction potential*, as a simple corollary for each finite value of $N \geq 2$.

Remark 7.1 Note that the dual norm (7.13) is a kind of weak norm. The error bound $\mathcal{O}(\Delta t)$ is consistent with the weak error estimate in [55].

7.2 Quantum Monte Carlo Methods

Computing the ground state energy of a many-body quantum system is a fundamental problem in chemistry. An important tool to determine the ground state energy and electron correlations is the quantum Monte Carlo (QMC) method [96, 4].

Consider the Hamiltonian,

$$\mathcal{H} = \sum_{i=1}^{N} -\frac{\hbar^2}{2m}\triangle_{x_i} + \sum_{i \neq j} W(x_i - x_j) + \sum_{i=1}^{N} V_{\text{ext}}(x_i). \tag{7.15}$$

Here V_{ext} is the external potential given by

$$V_{\text{ext}}(x_i) = \sum_{\alpha=1}^{M} U(x_i - R_\alpha), \tag{7.16}$$

where R_α, for instance, can be the position of an atom.

Up to some global phase factor, the ground state takes real values and is nonnegative everywhere. The ground state and the corresponding eigenvalue can be obtained via the Rayleigh quotient,

$$E = \min_{\Phi_N} \frac{\displaystyle\int_{(\mathbb{R}^3)^N} \Phi_N \mathcal{H} \Phi_N d\underline{x}}{\displaystyle\int_{(\mathbb{R}^3)^N} |\Phi_N|^2 d\underline{x}}, \tag{7.17}$$

where the minimizer Φ_N corresponds to the ground state wave function. The main computational challenge here is the curse of dimensionality due to the high dimensional integral.

In the variational Monte Carlo (VMC) framework, the ground state is approximated by selecting an appropriate ansatz $\Phi_N \approx \Phi_0$. Traditionally, Φ_0 is constructed using the one-body wave functions, by taking into the effect of particle correlations described by the Jastrow factors [32]. For example, in the Boson systems like the liquid Helium interacting with a graphite surface [80, 99, 86], the following ansatz has been proven successful,

$$\Phi_0 = e^{-J(\underline{x})}\Pi_{i=1}^N \phi(x_i), \quad J(\underline{x}) = \frac{1}{2}\sum_{i,j:i\neq j} u(|x_i - x_j|),$$

$$u(r) = \left(\frac{a}{r}\right)^5 + \frac{b^2}{r^2+c^2}. \tag{7.18}$$

The non-negative one-particle wave function is often taken as

$$\phi(x_i) = \sum_{\alpha=1}^{M} e^{-\theta(x_i - R_\alpha)}, \tag{7.19}$$

for some function θ. This form has been used in [99] and the parameters were obtained by solving a one-dimensional Schrödinger equation. With the approximation of Φ_N being fixed, the multi-dimensional integral is then interpreted as a statistical average. In fact, introducing the probability density function (PDF),

$$p(\underline{x}) \propto |\Phi_0(\underline{x})|^2, \tag{7.20}$$

the ground state energy is the average of E_{tot} under $p(\underline{x})$, where

$$E_{\text{tot}}(\underline{x}) = \frac{\mathcal{H}\Phi_0}{\Phi_0}. \tag{7.21}$$

Hence, E can be computed by a Monte Carlo procedure, and such a method is called the VMC, which is a typical QMC method.

In the VMC methods, the ground state is not updated. Instead, one may use another QMC method–the diffusion Monte Carlo (DMC) method [3, 87]–to compute the ground state and the energy. In particular, one solves a pseudo-time Schrödinger equation (TDSE) which is a parabolic equation [87]

$$\partial_t \Psi_N = (E_T - \mathcal{H}_N)\Psi_N. \tag{7.22}$$

Here, t represents a fictitious time. The energy shift E_T is adjusted on-the-fly based on the change of magnitude of the wave function. Instead of solving (7.22) directly, it is often more practical to find $f(\boldsymbol{r}, t)$ with

$$f(\underline{x}, t) = \Psi_N(\underline{x}, t)\Phi_0(\underline{x}). \tag{7.23}$$

By choosing $\Psi_N(\underline{x}, 0) = \Phi_0(\underline{x})$, $f(\underline{x}, 0) = |\Phi_0|^2 \propto p(\underline{x})$. Hence, a VMC method may be used to initialize $f(\underline{x}, t)$. Clearly, f solves the following differential equation [87],

$$\partial_t f = -\nabla \cdot \big(\frac{\hbar^2}{m} v(\underline{x}) f\big) + \frac{\hbar^2}{2m}\nabla^2 f - \big(E_T - E_{\text{tot}}(\underline{x})\big) f, \tag{7.24}$$

where $\underline{v} = (v_1, \cdots, v_N) \in \mathbb{R}^{Nd}$ and

$$v_i(\underline{x}) = \nabla \log \phi(x_i) - \sum_{j:j\neq i} \nabla_{x_i} u(|x_i - x_j|).$$

The average energy $E(t)$ is then defined as a weighted average,

$$E(t) = \frac{\displaystyle\int_{(\mathbb{R}^3)^N} f(\underline{x}, t) E_{\text{tot}}(\underline{x}) d\underline{x}}{\displaystyle\int_{(\mathbb{R}^3)^N} f(\underline{x}, t) d\underline{x}}, \tag{7.25}$$

where the correctness can be seen by $E(t) = \int \Psi_N \mathcal{H} \Phi_0 d\underline{x} \big/ \displaystyle\int \Psi_N \Phi_0 d\underline{x}$. If Ψ_N is close to the eigenstate, this will be close to E.

The key observation is that the dynamics (7.24) can be associated with a stochastic process, in which the particles are experiencing birth/death while driven by drift velocity and diffusion. This process can be implemented by a number of walkers together with birth/death processes [3, 87].

7.2.1 The Random Batch Method for VMC

With (7.18), the density (7.20) can be found as

$$p(\underline{x}) \propto e^{-2V}, \quad V = -\ln \Psi_0 = -\sum_i \log \phi(x_i) + \frac{1}{2} \sum_i \sum_{j\neq i} u(|x_i - x_j|), \tag{7.26}$$

and the total energy can be expressed as

$$E_{\text{tot}}(\underline{x}) = -\frac{\hbar^2}{2m} \triangle V - \frac{\hbar^2}{2m} \|\nabla V\|^2 + \sum_{i\neq j} W(x_i - x_j) + \sum_{i=1}^{N} \sum_{\alpha=1}^{M} U(x_i - R_\alpha). \tag{7.27}$$

To sample from $p(\underline{x})$, one may make use of the Markov chain Monte Carlo (MCMC) methods. Consider the overdamped Langevin dynamics,

$$d\boldsymbol{r}_i = \nabla \log \phi(\boldsymbol{r}_i) dt - \sum_{j:j\neq i} \nabla_{\boldsymbol{r}_i} u(|\boldsymbol{r}_i - \boldsymbol{r}_j|) dt + d\boldsymbol{W}_i(t), \quad 1 \leq i \leq N. \tag{7.28}$$

Under suitable conditions [78], the dynamical system with potential given by (7.26) is ergodic and the PDF $p(\underline{x})$ in (7.26) is the unique equilibrium measure of (7.28). By the classical Euler-Maruyama method [63], the underdamped Langevin can be discretized to a Markov Chain:

$$\begin{aligned} \boldsymbol{r}_i(t+\Delta t) &= \boldsymbol{r}_i(t) + \nabla \log \phi(\boldsymbol{r}_i)\Delta t \\ &\qquad - \sum_{j\neq i} \nabla_{\boldsymbol{r}_i} u(|\boldsymbol{r}_i(t) - \boldsymbol{r}_j(t)|)\Delta t + \Delta W_i, \quad 1 \le i \le N, \end{aligned} \tag{7.29}$$

where ΔW_i is again sampled from $\mathcal{N}(0, \Delta t)$. It is clear that $\mathcal{O}(M+N)$ operations should be taken for each particle at each time step.

The cost of the above MCMC is high. The strategy in [57] is to apply a RBM strategy with replacement. In particular, at each step, one randomly picks two particles, i and j, and compute their interactions, $\nabla_{\boldsymbol{r}_i} u(|\boldsymbol{r}_i - \boldsymbol{r}_j|)$, then updates their positions as follows,

$$\begin{cases} \boldsymbol{r}_i(t+\Delta t) = \boldsymbol{r}_i(t) + \nabla \log \phi(\boldsymbol{r}_i)\Delta t + (N-1)\nabla_{\boldsymbol{r}_i} u(|\boldsymbol{r}_i - \boldsymbol{r}_j|)\Delta t + \Delta W_i, \\ \boldsymbol{r}_j(t+\Delta t) = \boldsymbol{r}_j(t) + \nabla \log \phi(\boldsymbol{r}_j)\Delta t + (N-1)\nabla_{\boldsymbol{r}_j} u(|\boldsymbol{r}_i - \boldsymbol{r}_j|)\Delta t + \Delta W_j. \end{cases} \tag{7.30}$$

For the one-body term $\nabla \log \phi(\boldsymbol{r}_i)$,

$$\nabla \log \phi(\boldsymbol{r}_i) = \sum_{\alpha=1}^{M} -\nabla\theta(\boldsymbol{r}_i - R_\alpha) q_\alpha^i, \quad q_\alpha^i = \frac{e^{-\theta(\boldsymbol{r}_i - R_\alpha)}}{\sum_{\beta=1}^{M} e^{-\theta(\boldsymbol{r}_i - R_\beta)}}, \tag{7.31}$$

where the coefficients q_α^i's are non-negative and $\sum_\alpha q_\alpha^i = 1$. To reduce the cost, one may further use a direct Monte Carlo method: pick *just one* term α randomly. Specifically, assume that one starts with α and computes $e_{\text{old}} = \theta(\boldsymbol{r}_i - R_\alpha)$, and then one randomly picks $1 \le \beta \le M$ and computes $e_{\text{new}} = \theta(\boldsymbol{r}_i - R_\beta)$. β is accepted with probability

$$p_{\text{acc}} \propto \exp\big[-(e_{\text{new}} - e_{\text{old}})\big]. \tag{7.32}$$

For the detailed algorithm see [57]. As a result of the random sampling of the one- and two-body interactions, updating the position of each particle *only requires* $\mathcal{O}(1)$ *operations* per time step. Another practical issue emerges when the interaction $u(|x|)$ has a singularity near zero. One can use the splitting idea as mentioned in Sect. 4.1, i.e., applying RBM only to the long-range smooth part.

It was shown in [57] that the above random batch algorithm, when applied to one batch of two particles, has the same accuracy as the Euler-Maruyama method over a time step of $2\Delta t/N$. One full time step in Euler-Maruyama method corresponds to $N/2$ such steps in the random batch algorithm. This corresponds to the random batch method with replacement.

We show a numerical experiment performed in [57] on ^{4}He atoms interacting with a two-dimensional lattice. The CPU times taken to move the 300 Markov chains for 1000 steps were compared. In this comparison, the cost associated with the energy calculations was excluded in the random batch and Euler-Maruyama methods. From Table 3, one clearly sees that the RBM is more efficient than the Euler-Maruyama method. It is much more efficient than the random walk

Table 3 CPU times (seconds) for several VMC methods

	Random walk metropolis-hastings	Euler-Maruyama	Random batch
CPU time for a 1000-step sampling period	1503	469	54

Metropolis-Hastings algorithm, mainly because the latter method requires the calculation of the energy at *every* step.

7.2.2 The Random Batch Method for DMC

Viewing (7.24), one may consider an ensemble of L copies of the system, also known as walkers [3]. For each realization, one first solves the SDEs corresponding to the drift and diffusion, which is the same as the overdamped Langevin as in VMC up to a time scaling. Hence, the same Random Batch Algorithm in the VMC can be used for this part.

The relaxation term $-(E_T - E_{\text{tot}})f$ is then done by using a birth/death process to determine whether a realization should be removed or duplicated. For each walker, one computes a weight factor,

$$w(t+\Delta t) = \exp\Big[\Delta t\big(E_T - \tfrac{1}{2}(E_{\text{tot}}(\boldsymbol{r}) + E_{\text{tot}}(\boldsymbol{r}'))\big)\Big]. \tag{7.33}$$

This weight determines how the walker should be removed or duplicated. See [57] for more details. The primary challenge is that computing the energy at each step requires $\mathcal{O}((N+M)N)$ operations in order to update the position of N particles. To reduce this part of the computation cost, one rewrites the total energy as

$$E_{\text{tot}}(\boldsymbol{r}) = \sum_{i=1}^{N} E_1(\boldsymbol{r}_i) + \sum_{1\le i<j\le N} E_2(\boldsymbol{r}_i, \boldsymbol{r}_j) + \sum_{1\le i<j<k\le N} E_3(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k), \tag{7.34}$$

where

$$\begin{aligned} E_1(\boldsymbol{r}_i) &= -\frac{\hbar^2}{2m}\nabla^2 \ln\phi(\boldsymbol{r}_i) - \frac{\hbar^2}{2m}|\nabla \ln\phi(\boldsymbol{r}_i)|^2 + \sum_{\alpha=1}^{M} U(\boldsymbol{r}_i - R_\alpha), \\ E_2(\boldsymbol{r}_i, \boldsymbol{r}_j) &= -\frac{\hbar^2}{m}\nabla^2 \ln u(r_{ij}) + \frac{\hbar^2}{m}\big(\nabla \ln\phi(\boldsymbol{r}_i) - \nabla \ln\phi(\boldsymbol{r}_j)\big)\cdot \nabla u(r_{ij}) \\ &\quad + \frac{\hbar^2}{m}|\nabla u(r_{ij})|^2 + W(r_{ij}), \end{aligned}$$

$$E_3(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k) = \frac{\hbar^2}{m}\Big[\nabla u(r_{ij}) \cdot \nabla u(r_{ik}) + \nabla u(r_{ji}) \cdot \nabla u(r_{jk}) + \nabla u(r_{ki}) \cdot \nabla u(r_{kj})\Big]. \tag{7.35}$$

Here, $\boldsymbol{r}_{ij} = \boldsymbol{r}_i - \boldsymbol{r}_j$ and $r_{ij} = |\boldsymbol{r}_{ij}|$. The three-body terms arise because of the $\|\nabla V\|^2$ term in (7.27).

In the random batch algorithm proposed in [57], one randomly picks a batch C_I with three particles: $C_I = \{i, j, k\}$. One first updates the position of the three particles (drift and diffusion) by solving the overdamped Langevin dynamics using the random batch algorithm with batch size 3. Then, one then defines a *local* energy,

$$\begin{aligned} E_I(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k) =& E_1(\boldsymbol{r}_i) + E_1(\boldsymbol{r}_j) + E_1(\boldsymbol{r}_k) \\ &+ \tfrac{N-1}{2}\Big[E_2(\boldsymbol{r}_i, \boldsymbol{r}_j) + E_2(\boldsymbol{r}_j, \boldsymbol{r}_k) + E_2(\boldsymbol{r}_k, \boldsymbol{r}_i)\Big], \\ &+ \tfrac{(N-1)(N-2)}{2} E_3(\boldsymbol{r}_i, \boldsymbol{r}_j, \boldsymbol{r}_k), \end{aligned} \tag{7.36}$$

where in E_1, the sum $\sum_{\alpha=1}^{M}$ can be further reduced by a mini-batch strategy. Computing this local energy is clearly $\mathcal{O}(1)$. To avoid frequent removal and duplication of walkers, the branching process is applied after $N/3$ batches of particles are updated. In this case, the weight function is defined by collecting the local energy from each batch (denoted by I_m here),

$$w(\boldsymbol{r}) = \exp\big[\Delta t\big(E_T - \widetilde{E}_{\text{tot}}\big)\big], \quad \widetilde{E}_{\text{tot}} = \sum_{m=1}^{N/3} E_{I_m}. \tag{7.37}$$

Because of the smallness of Δt, the expectation of w_I equals $w(t + \Delta t)$ modulus an error of $\mathcal{O}(\Delta t^2)$. See [57] for the verification using the Green's functions.

The detailed algorithm can be found in [57] and we omit it here. Now we show a test of the RBM-DMC algorithm conducted in [57], which compares the results with the direct DMC method. For the initialization, a VMC method using the ansatz (7.18) for the wave function Φ_0 was first applied. The random walk Metropolis-Hastings Monte Carlo method is used in both methods so that they start at the same states. 300 ensembles are created by sub-sampling one sample out of every 500 steps from the VMC runs to avoid correlations among the ensembles. For both methods, $\Delta t = 10^{-4}$ was used and 200,000 steps of simulations were run. The CPU runtime is recorded for various system sizes. More specifically, the system size is increased from the original 168 particles, to $N = 378$, $N = 672$, and $N = 1050$ particles, and in each case, the direct DMC and the RBM-DMC were run for 1000 steps. As shown in Fig. 8, the CPU time for the direct DMC method increases much more rapidly as N increases.

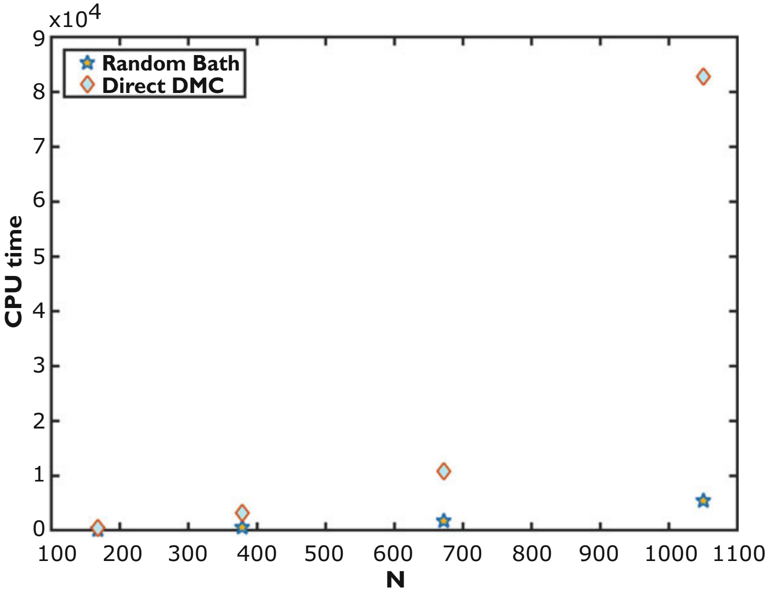


Fig. 8 A comparison of the CPU runtime (in seconds) for running 1000 steps of DMC

Acknowledgments S. Jin was partially supported by the NSFC grant No.12031013. The work of L. Li was partially sponsored by NSFC 11901389, 11971314, and Shanghai Sailing Program 19YF1421300. Both authors were also supported by Shanghai Science and Technology Commission Grant No. 20JC144100.

References

1. G. Albi, N. Bellomo, L. Fermo, S-Y Ha, J. Kim, L. Pareschi, D. Poyato, and J. Soler. Vehicular traffic, crowds, and swarms: From kinetic theory and multiscale methods to applications and research perspectives. *Mathematical Models and Methods in Applied Sciences*, 29(10):1901–2005, 2019.
2. G. Albi and L. Pareschi. Binary interaction algorithms for the simulation of flocking and swarming dynamics. *Multiscale Modeling & Simulation*, 11(1):1–29, 2013.
3. J. B. Anderson. A random-walk simulation of the Schrödinger equation: H+3. *The Journal of Chemical Physics*, 63(4):1499–1503, 1975.
4. J. B. Anderson. *Quantum Monte Carlo: origins, development, applications*. Oxford University Press, 2007.
5. H. Babovsky and R. Illner. A convergence proof for Nanbu's simulation method for the full Boltzmann equation. *SIAM journal on numerical analysis*, 26(1):45–65, 1989.
6. W. Bao, S. Jin, and P. A. Markowich. On time-splitting spectral approximations for the Schrödinger equation in the semiclassical regime. *Journal of Computational Physics*, 175(2):487–524, 2002.
7. J. Barnes and P. Hut. A hierarchical O(NlogN) force-calculation algorithm. *Nature*, 324:446–449, 1986.

8. A. L. Bertozzi, J. B. Garnett, and T. Laurent. Characterization of radially symmetric finite time blowup in multidimensional aggregation equations. *SIAM J. Math. Anal.*, 44(2):651–681, 2012.
9. U. Biccari and E. Zuazua. A stochastic approach to the synchronization of coupled oscillators. *Front. Energy Res.*, 8(115), 2020.
10. G. A. Bird. Approach to translational equilibrium in a rigid sphere gas. *The Physics of Fluids*, 6(10):1518–1519, 1963.
11. L. Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.
12. George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.
13. S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3–4):231–357, 2015.
14. H. B. Callen and T. A. Welton. Irreversibility and generalized noise. *Physical Review*, 83(1):34, 1951.
15. E. Carlen, P. Degond, and B. Wennberg. Kinetic limits for pair-interaction driven master equations and biological swarm models. *Mathematical Models and Methods in Applied Sciences*, 23(07):1339–1376, 2013.
16. J. A. Carrillo, L. Pareschi, and M. Zanella. Particle based gPC methods for mean-field models of swarming with uncertainty. *Communications in Computational Physics*, 25(2), 2019.
17. Y.-P. Choi, S.-Y. Ha, and S.-B. Yun. Complete synchronization of Kuramoto oscillators with finite inertia. *Physica D: Nonlinear Phenomena*, 240(1):32–44, 2011.
18. G. Ciccotti, D. Frenkel, and I. R. McDonald. *Simulation of liquids and solids: Molecular Dynamics and Monte Carlo Methods in Statistical Mechanics*. North-Holland, Amsterdam, 1987.
19. F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Transactions on automatic control*, 52(5):852–862, 2007.
20. B. Dai, N. He, H. Dai, and L. Song. Provable Bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994, 2016.
21. P. Degond, J.-G. Liu, and R. L. Pego. Coagulation–fragmentation model for animal group-size statistics. *Journal of Nonlinear Science*, 27(2):379–424, 2017.
22. P. Degond, J.-G. Liu, and C. Ringhofer. Evolution of the distribution of wealth in an economic environment driven by local Nash equilibria. *Journal of Statistical Physics*, 154(3):751–780, 2014.
23. Markus Deserno and Christian Holm. How to mesh up Ewald sums. II. An accurate error estimate for the particle-particle particle-mesh algorithm. *The Journal of Chemical Physics*, 109(18):7694–7701, 1998.
24. G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl. A Stein variational Newton method. In *Advances in Neural Information Processing Systems*, pages 9187–9197, 2018.
25. Z. H. Duan and R. Krasny. An Ewald summation based multipole method. *J. Chem. Phys.*, 113:3492–3495, 2000.
26. J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
27. R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 4 edition, 2010.
28. R. Durstenfeld. Algorithm 235: random permutation. *Communications of the ACM*, 7(7):420, 1964.
29. Weinan E, Tiejun Li, and Eric Vanden-Eijnden. *Applied stochastic analysis*, volume 199. American Mathematical Soc., 2019.
30. A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.
31. L. Erdos and H.-T. Yau. Dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics*, 28, 2017.

32. WMC Foulkes, Lubos Mitas, RJ Needs, and Guna Rajagopal. Quantum Monte Carlo simulations of solids. *Reviews of Modern Physics*, 73(1):33, 2001.
33. R. H. French, V. A. Parsegian, R. Podgornik, R. F. Rajter, A. Jagota, J. Luo, D. Asthagiri, M. K. Chaudhury, Y.-M. Chiang, S. Granick, S. Kalinin, M. Kardar, R. Kjellander, D. C. Langreth, J. Lewis, S. Lustig, D. Wesolowski, J. S. Wettlaufer, W.-Y. Ching, M. Finnis, F. Houlihan, O. A. von Lilienfeld, C. J. van Oss, and T. Zemb. Long range interactions in nanoscale science. *Rev. Mod. Phys.*, 82(2):1887–1944, 2010.
34. D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.
35. D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.
36. Y. Gao and J.-G. Liu. A note on parametric Bayesian inference via gradient flows. *Annals of Mathematical Sciences and Applications*, 5(2):261–282, 2020.
37. A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Reviews of Modern Physics*, 68(1):13, 1996.
38. S. Gershman, M. Hoffman, and D. Blei. Nonparametric variational inference. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 235–242, 2012.
39. W. R Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC, 1995.
40. F. Golse. The mean-field limit for the dynamics of large particle systems. *Journées équations aux dérivées partielles*, 9:1–47, 2003.
41. F. Golse, S. Jin, and T. Paul. The random batch method for n-body quantum dynamics. *J. Comp. Math.*, arXiv preprint arXiv:1912.07424 (2019).
42. L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73:325–348, 1987.
43. S. Y. Ha, S. Jin, D. Kim, and D. Ko. Convergence toward equilibrium of the first-order consensus model with random batch interactions. Journal of Differential Equations, 302, 585–616, 2021.
44. S.-Y. Ha and Z. Li. Complete synchronization of Kuramoto oscillators with hierarchical leadership. *Communications in Mathematical Sciences*, 12(3):485–508, 2014.
45. S.-Y. Ha and J.-G. Liu. A simple proof of the Cucker-Smale flocking dynamics and mean-field limit. *Commun. Math. Sci.*, 7(2):297–325, 2009.
46. Seung-Yeal Ha, Shi Jin, Doheon Kim, and Dongnam Ko. Uniform-in-time error estimate of the random batch method for the Cucker-Smale model. *Math. Models Methods Appl. Sci.*, 31(6):1099–1135, 2021.
47. Seung-Yeal Ha and Eitan Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *Kinet. Relat. Models*, 1(3):415–435, 2008.
48. W. K. Hastings. *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*. Oxford University Press, 1970.
49. B. Hetenyi, K. Bernacki, and B. J. Berne. Multiple "time step" Monte Carlo. *J. Chem. Phys.*, 117(18):8203–8207, 2002.
50. W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.
51. D. Horstmann. From 1970 until present: the Keller-Segel model in chemotaxis and its consequences. *Jahresber. Dtsch. Math.-Ver.*, 105:103–165, 2003.
52. Pierre-Emmanuel Jabin and Zhenfu Wang. Mean field limit for stochastic particle systems. In *Active Particles, Volume 1*, pages 379–402. Springer, 2017.
53. S. Jin and L. Li. On the mean field limit of the Random Batch Method for interacting particle systems. *Science China Mathematics*, pages 1–34, 2021.
54. S. Jin, L. Li, and J.-G. Liu. Random Batch methods (RBM) for interacting particle systems. *Journal of Computational Physics*, 400:108877, 2020.

55. S. Jin, L. Li, and J.-G. Liu. Convergence of the random batch method for interacting particles with disparate species and weights. *SIAM Journal on Numerical Analysis*, 59(2):746–768, 2021.
56. S. Jin, L. Li, and Y. Sun. On the Random Batch Method for second order interacting particle systems. *arXiv preprint arXiv:2011.10778*, 2020.
57. S. Jin and X. Li. Random batch algorithms for quantum Monte Carlo simulations. *Commun. Comput. Phys.*, 28(5):1907–1936, 2020.
58. S. Jin, P. Markowich, and C. Sparber. Mathematical and computational methods for semiclassical Schrödinger equations. *Acta Numerica*, 20:121–209, 2011.
59. Shi Jin, Lei Li, Zhenli Xu, and Yue Zhao. A Random Batch Ewald Method for Particle Systems with Coulomb Interactions. *SIAM J. Sci. Comput.*, 43(4):B937–B960, 2021.
60. J. K. Johnson, J. A. Zollweg, and K. E. Gubbins. The Lennard-Jones equation of state revisited. *Molecular Physics*, 78(3):591–618, 1993.
61. M. H. Kalos and P. A. Whitlock. *Monte Carlo methods*. John Wiley & Sons, 2009.
62. K. Kawasaki. Simple derivations of generalized linear and nonlinear Langevin equations. *Journal of Physics A: Mathematical, Nuclear and General*, 6(9):1289, 1973.
63. P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media, 2013.
64. D. Ko, S.-Y. Ha, S. Jin, and D. Kim. Uniform error estimates for the random batch method to the first-order consensus models with antisymmetric interaction kernels. *Studies Appl. Math.*, 146(4):983–1022, 2021.
65. D. Ko and Z. Enrique. Model predictive control with random batch methods for a guiding problem. *Mathematical Models and Methods in Applied Sciences*, 31(8):1569-1592, 2021.
66. J.-M. Lasry and P.-L. Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.
67. L. Li, Y. Li, J.-G. Liu, Z. Liu, and J. Lu. A stochastic version of Stein variational gradient descent for efficient sampling. *Communications in Applied Mathematics and Computational Science*, 15(1):37–63, 2020.
68. L. Li, J.-G. Liu, and Y. Tang. A direct simulation approach for the Poisson-Boltzmann equation using the Random Batch Method. *arXiv preprint arXiv:2004.05614*, 2020.
69. L. Li, J.-G. Liu, and P. Yu. On mean field limit for Brownian particles with Coulomb interaction in 3D. *J. Math. Phys.*, 60(111501), 2019.
70. L. Li, Z. Xu, and Y. Zhao. A random-batch Monte Carlo method for many-body systems with singular kernels. *SIAM Journal on Scientific Computing*, 42(3):A1486–A1509, 2020.
71. J. Liang, P. Tan, Y. Zhao, L. Li, S., Jin, L. Hong, and Z. Xu. Superscalability of the random batch Ewald method. *J. Chem. Phys.*, 156, 014114 (2022).
72. Evgenii Mikhailovich Lifshitz and Lev Petrovich Pitaevskii. *Statistical physics: theory of the condensed state*, volume 9. Elsevier, 2013.
73. Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.
74. Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.
75. J. Lu, Y. Lu, and J. Nolen. Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM J. Math. Anal.*, 51(2):648–671, 2019.
76. B. A. Luty, M. E. Davis, I. G. Tironi, and W. F. Van Gunsteren. A comparison of particle-particle, particle-mesh and Ewald methods for calculating electrostatic interactions in periodic molecular systems. *Mol. Simul.*, 14:11–20, 1994.
77. M. G. Martin, B. Chen, and J. I. Siepmann. A novel Monte Carlo algorithm for polarizable force fields: application to a fluctuating charge model for water. *The Journal of chemical physics*, 108(9):3383–3385, 1998.
78. J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

79. H. P. McKean. Propagation of chaos for a class of non-linear parabolic equations. *Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967)*, pages 41–57, 1967.
80. W. L. McMillan. Ground state of liquid he4. *Physical Review*, 138(2A):A442, 1965.
81. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.
82. G. N. Milstein and M. V. Tretyakov. *Stochastic numerics for mathematical physics*. Springer Science & Business Media, 2013.
83. S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.
84. K. Nanbu. Direct simulation scheme derived from the Boltzmann equation. i. monocomponent gases. *Journal of the Physical Society of Japan*, 49(5):2042–2049, 1980.
85. S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics*, 52(2):255–268, 1984.
86. T. Pang. Diffusion Monte Carlo: a powerful tool for studying quantum many-body systems. *American Journal of Physics*, 82(10):980–988, 2014.
87. P. J. Reynolds, D. M. Ceperley, B. J. Alder, and W. A. Lester Jr. Fixed-node quantum Monte Carlo for molecules. *The Journal of Chemical Physics*, 77(11):5593–5603, 1982.
88. D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.
89. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.
90. V. Rokhlin. Rapid solution of integral equations of classical potential theory. *Journal of computational physics*, 60(2):187–207, 1985.
91. F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.
92. H. E. Stanley. *Phase transitions and critical phenomena*. Clarendon Press, Oxford, 1971.
93. Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.
94. J. Toner and Y. Tu. Flocks, herds, and schools: A quantitative theory of flocking. *Physical review E*, 58(4):4828, 1998.
95. T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.
96. W. von der Linden. A quantum Monte Carlo approach to many-body physics. *Physics Reports*, 220(2–3):53–162, 1992.
97. R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686, 2019.
98. M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.
99. PA Whitlock, GV Chester, and B Krishnamachari. Monte Carlo simulation of a helium film on graphite. *Physical Review B*, 58(13):8704, 1998.
100. A. T. Winfree. *The geometry of biological time*, volume 12. Springer Science & Business Media, 2001.
101. L. Ying, G. Biros, and D. Zorin. A kernel-independent adaptive fast multipole algorithm in two and three dimensions. *J. Comput. Phys.*, 196:591–626, 2004.

Trends in Consensus-Based Optimization

Claudia Totzeck

Abstract In this chapter we give an overview of the consensus-based global optimization algorithm and its recent variants. We recall the formulation and analytical results of the original model, and then we discuss variants using component-wise independent or common noise. In combination with mini-batch approaches those variants were tailored for machine learning applications. Moreover, it turns out that the analytical estimates are dimension independent, which is useful for high-dimensional problems. We discuss the relationship of consensus-based optimization with particle swarm optimization, a method widely used in the engineering community. Then we survey a variant of consensus-based optimization that is proposed for global optimization problems constrained to hyper-surfaces. We conclude the chapter with remarks on applications, preprints and open problems.

1 Introduction

Global optimization tasks arise in various fields such as economics, finance, physics, clustering and artificial intelligence. In the most general form, these read

$$\min_{x \in \mathcal{X}} f(x)$$

for a given objective function f and state space $\mathcal{X}$. Despite its simple description, the problem is nontrivial for nonconvex f with possibly many local minima or constraint state spaces $\mathcal{X}$; see Fig. 1. Its importance in various disciplines attracted the attention of many researchers to seek for solution strategies. Here, we focus on agent-based methods: on the one hand, there are biologically inspired methods as the ant colony optimization [1], artificial bee colony optimization [2] or firefly optimization [3].

C. Totzeck (✉)
School of Mathematics and Natural Sciences, University of Wuppertal, Wuppertal, Germany
e-mail: totzeck@uni-wuppertal.de

N. Bellomo et al. (eds.), *Active Particles, Volume 3*, Modeling and Simulation in Science, Engineering and Technology, https://doi.org/10.1007/978-3-030-93302-9_6

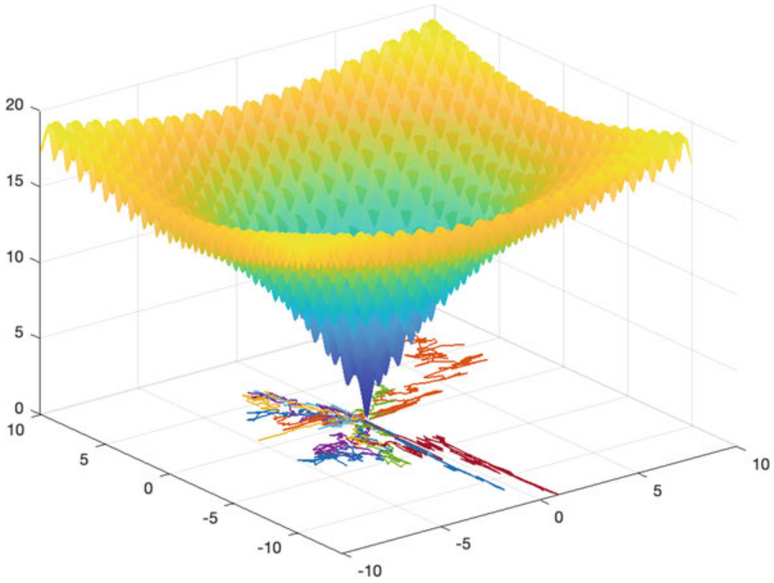


Fig. 1 Plot of the Ackley [4] benchmark function for global optimization in two dimensions with trajectories of one realisation of (6) with 20 particles visualized in the xy-plane

On the other hand, wind-driven optimization (WDO) [5] is physically inspired as it models weather phenomena such as pressure and wind. The most popular agent-based global optimization algorithms are the Particle Swarm Optimization (PSO) and Simulated Annealing (SA). In PSO agents explore the state space while encountering a randomized drift towards the global best position seen by all the agents and a second drift towards their personal best positions. We will see more details on PSO below in Sect. 3 where similarities and differences of CBO and PSO are discussed. SA is physically inspired, again, agents explore the state space. They are driven by noise terms that are diminishing as time evolves. The decrease of stochastic influence is called cool down and the particles are expected to concentrate at the best position seen by the particles during the exploration phase.

Most of the global optimization approaches are heuristics that have proven to give useful results in applications but lack a rigorous analysis. Some proofs of convergence exist for SA. The ones in the context of image restoration and global optimization are mostly in the discrete setting and based on Markov chains; see the survey [6] for more details. Another proof considers a kinetic formulation of SA [7]. It was then generalized to Langevin-based SA in [8].

A main objective in the modelling of the CBO scheme was to treat all particles identically, in particular, to circumvent the selection of a current best particle. In this way, one expects to have a corresponding mean-field scheme that can be utilized for the convergence analysis. Having this in mind, the CBO method is proposed as a system of stochastic differential equation (SDE) that mimics interacting agents communicating over a weighted mean. By construction, the particles are expected to build a consensus at the position of the weighted mean that is located near the global minimizer of the functional.

To achieve this behaviour, CBO combines ideas of swarm intelligence [9] with approaches from consensus formation [10] in order to obtain a scheme that minimizes the objective function. CBO was first introduced in [11], where formal relations to the mean-field equation and promising numerical results were shown. The main feature of the CBO algorithm is a weighted mean, v_f. Particles with small function values have more influence in the weighted mean than particles with large function values. In this way the weighted mean is expected to be a good approximation of the global minimizer. All particles are driven by two terms. A drift term forcing them to move towards the weighted mean and a scaled diffusion allowing for exploration. In fact, whenever a particle is far away from the weighted mean, it explores its surroundings and tries to find a better position than the weighted average has. The scaling of the diffusion depends on the distance of the particle to the weighted mean. If the two coincide, the diffusion vanishes. Hence, the scheme allows for concentration at the position of the weighted mean.

The fact that the global minimizer is approximated with the help of the weighted mean is crucial when it comes to the mean-field limit. In particular, using the weighted mean the scheme circumvents to label any particle as current leader, or current global best, which would make the particles distinguishable and prevent us to carry out the mean-field limit. Formally, the limiting equation for 'number of particles to infinity' is the PDE corresponding to the McKean process resulting from Itô calculus applied to the SDE system [11, 12]. In [12] a rigorous analysis of the PDE method is performed. In particular, it is shown that the method converges to the minimizer of the global optimization scheme under some appropriate conditions.

Another advantage of the communication via the weighed mean is a reduction of the computational effort. In fact, the communication with the weighted mean is of order $\mathcal{O}(N)$ for N particles in the swarm. In other consensus algorithms each particle communicates with all other particle separately, leading to an effort of $\mathcal{O}(N^2)$, which suffers the curse of dimensionality when the swarm size grows.

Recently, variants and extensions were proposed to improve the CBO method. Some approaches aim to enhance the performance in high-dimensional problems such as the ones arising in machine learning. Others extend the class of problems to be solved with CBO, for example, they allow for constrained state space $\mathcal{X}$.

In this survey we discuss these advances and compare them to the original method. The main part covers models that have been approved by peer review. At the end, we shed some light on recent preprints as well. Before we go into the details, we shortly describe the ideas covered in the following.

In [13] the diffusion term was replaced by a component-wise diffusion, leading to a scheme that is robust with respect to the dimension of the state space. Indeed, the authors were able to show that many of the estimates shown in [12] hold without dimension dependence for the scheme with component-wise diffusion. Moreover, the article introduces a mini-batch idea for the computation of the weighted mean. This reduces the computational cost and has positive effects on the performance in high-dimensional scenarios. More details on this variant are discussed in Sect. 2.2. The authors of [14] replace the component-wise independent noise of the above variant by a component-wise common noise. This adaption facilitates the analysis of the scheme on the particle level. In fact, the authors show convergence of the variant directly on the particle level in contrast to [12, 13], where the PDE formulation is employed for the analysis. A variant that incorporates global in time information in order to approximate the personal best position seen by each of the particles is proposed in [15]. It is shown that this variant is robust even if the initial distribution of particles is inconvenient. We discuss the scheme with global in time information and its relationship to PSO in Sect. 3.

In addition, there are variants that take care of optimization problems on constrained sets. Box constraints are rather simple to handle. Dynamics constrained to hyper-surfaces, for example the sphere, need more sophisticated ideas [16]. We discuss approaches for constraint sets in Sect. 4, and in Sect. 5 we briefly comment on the performance of the CBO variants. For example in [13] are comparisons to stochastic gradient descent (SGD) methods and several studies for global optimization benchmarks reported. We conclude with an outlook to future work and open problems.

1.1 Notation and Assumptions

Let us first fix the notation and assumptions that are consistently used in the following sections. This has the advantage that the sections are self-consistent and one might jump to the variant of most interest right after the introduction.

We denote the dimension of the state space by $d \in \mathbb{N}$ and $N \in \mathbb{N}$ is the number of agents or particles in the swarm. The two notions, agents and particles, are used equally throughout the text. The state of the i-th agent is given by a vector $X^i \colon [0,T] \to \mathbb{R}^d, i = 1, \dots, N,$, and we collect the states of all agents at time $t \in [0,T]$ in the vector $X(t) = (X^1(t), \dots, X^N(t)) \in \mathbb{R}^{dN}$. The initial condition of the particles is denoted by $X_0^i \in \mathbb{R}^d$ for $i = 1, \dots, N$, and we assume that X_0^i are independent and identically distributed with $\text{law}(X_0^i) = \rho_0 \in \mathcal{P}(\mathbb{R}^d)$. The constants $\lambda, \sigma \geq 0$ denote the drift and diffusion parameters, respectively. Some schemes incorporate a Heaviside function H or a regularization H^ϵ thereof, which we fix as

$$H(x) = \begin{cases} 1, & \text{for } x \geq 0, \\ 0, & \text{else} \end{cases}, \qquad H^\epsilon(x) = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{x}{\epsilon}\right).$$

Moreover, we denote by $W^i, i = 1, \dots, N$, independent d-dimensional Brownian motions. We consider the minimization problem

$$\min_{x \in \mathcal{X}} f(x), \tag{P}$$

where $f: \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a continuous function that admits a unique global minimizer $X_* \in \mathbb{R}^d$ and $\mathcal{X} = \mathbb{R}^d$ except for Sect. 4.1, where we discuss state constraints and minimize f on some hyper-surface $\Gamma \subset \mathbb{R}^d$.

1.1.1 The Weighted Average

As mentioned above, a weighted average or weighted mean plays a crucial role in all variants of CBO. For simplicity, we fix the weight function to be

$$\omega_\alpha^f(x) = \exp(-\alpha f(x)) \tag{1}$$

throughout this review. Other choices are possible as well, but the weight function should be tailored to represent the task of finding a global minimum.

Unless otherwise stated, the notion *weighted mean* refers to the vector

$$v_f = \frac{1}{\sum_{i=1}^N \omega_\alpha^f(X^i(t))} \sum_{i=1}^N X^i(t)\omega_\alpha^f(X^i(t)). \tag{2}$$

Note that the objective function enters into the weight. Hence, due to (1) agents at locations with lower function values have more weight in the mean than agents located at positions with high function values. The parameter α controls this separation effect. Indeed, for $\alpha = 0$ all particles have the same weight and for $\alpha \to \infty$ we expect v_f to approximate the global best of the agents, i.e.,

$$v_f \approx \operatorname{argmin}_{i=1,\dots,N} f(X^i(t)).$$

Note that the argmin may be set-valued in general. For simplicity, we assumed above that f attains a unique minimizer.

The argument for $\alpha \to \infty$ is related to the Laplace principle from large deviation theory [17]. In fact, under the assumption that the processes $X^i(t)$ are independent, we formally pass to the limit $N \to \infty$ to obtain

$$\frac{1}{\sum_{i=1}^N \omega_\alpha^f(X^i(t))} \sum_{i=1}^N X^i(t)\omega_\alpha^f(X^i(t)) \to \frac{1}{\int \omega_\alpha^f(x)d\rho_t} \int x\omega_\alpha^f(x)d\rho_t$$

in distributional sense, with $\rho_t \in \mathcal{P}^{ac}(\mathbb{R}^d)$ being the Borel probability measure describing the one-particle mean-field distribution. Here $\mathcal{P}^{ac}(\mathbb{R}^d)$ denotes the space of Borel probability measures that are absolutely continuous w.r.t. the Lebesgue measure dx. Then, by Laplace principle [11] we have

Proposition 1 *Assume that $f \in C_b(\mathbb{R}^d, \mathbb{R})$, $f \geq 0$, attains a unique global minimum at the point $X_* \in \mathbb{R}^d$, and let $\rho \in \mathcal{P}^{ac}(\mathbb{R}^d)$. Then, it holds*

$$\lim_{\alpha\to\infty}\left(-\frac{1}{\alpha}\log\left(\int_{\mathbb{R}^d} e^{-\alpha f}\,d\rho\right)\right) = f(X_*).$$

This property is the main motivation to choose the ω_α^f as given in (1); see [11, Proposition 2.1] for details. Note that uniqueness of the minimizer plays a role. If there were several global minimizers, the weighted mean would be in the convex hull of these and, in general, have a greater function value.

In the following section we recall the original statement of the CBO scheme, then we discuss recent variants. Readers familiar with the original scheme may jump directly to the variant of their interest.

2 Consensus-Based Global Optimization Methods

We begin this section with the original method as proposed in [11] and analysed in [12]. Then we move on to recent variants that were tailored to improve the method for high-dimensional applications as arising in machine learning. The variants replace the diffusion term with either component-wise independent or component-wise common diffusion.

2.1 Original Statement of the Method

The ideas behind and main features of CBO [11] are explained on the particle level. Then we formally pass to the mean-field level and review analytical results that discuss the formation of consensus near the global minimizer [12].

2.1.1 Particle Scheme

Consensus-based optimization was first introduced in [11] as a swarm dynamic that consists of N coupled stochastic differential equations (SDEs). The equation of the i-th agent is given by

$$dX^i(t) = -\lambda(X^i(t)-v_f)H^\epsilon(f(X^i(t))-f(v_f))dt + \sqrt{2}\sigma|X^i(t)-v_f|dW^i(t), \tag{3}$$

for $i = 1,\dots,N$ and supplemented with initial data $X(0) = X_0$. The system is coupled by the weighted average, v_f, which appears in the equation of every agent. The first term on the right-hand side models a drift towards v_f. The greater the distance of the agent's position to v_f, the stronger the drift. The Heaviside function assures that the particle only moves towards v_f, if the function value of v_f is better, i.e. smaller than the function value of the particle. The idea behind the diffusion term is similar. The diffusivity is scaled with the distance of $|X^i(t) - v_f|$, an agent far away from v_f is allowed to explore its neighbourhood and possibly find a better position than v_f. While an agent close to v_f is less diffusive and tends to keep its position. In particular, the diffusion of particle i vanishes if $X^i = v_f$. This allows for concentration of the particles at v_f.

Remark 1 Let us emphasize some advantages of this dynamic:

1. **Indistinguishable particles:** Compared to other swarm intelligence schemes the dynamic does not depend on $\text{argmin}_{X^i} f(X^i)$, but only on its approximation v_f. Therefore, we may formally derive a limiting equation in mean-field sense as $N \to \infty$, compare Sect. 2.1.2, and use the PDE for the analytical investigation.
2. **Interaction scales with N:** The coupling via v_f has a huge advantage as well from a numerical point of view as we do not have binary interactions. The effort for the interaction of the agents scales only linearly in N. This is in contrast to many interaction models for crowd dynamics, where agents interact with all other agents at the same time, leading to a convolution term of order $\mathcal{O}(N^2)$.
3. **Exploration of full space:** Due to the term $|X^i(t)-v_f|dW^i(t)$, exploration takes place in $\mathbb{R}^d$ even if the X^i are initially spanning only a subspace of $\mathbb{R}^d$. This has a positive effect on the exploration if $N \ll d$.

Heaviside Function

In the original model, the Heaviside function was imposed to make concentration in local minima less probable. As reported in [11], the deterministic scheme, $\sigma = 0$, with Heaviside function allows for stationary solutions consisting of several Dirac measures located at level sets of f. For $\sigma > 0$, these solutions have probability zero, due to the Brownian motion. Moreover, it turned out in numerical studies that the scheme works fine without the multiplication of the Heaviside function, and for the analytical investigation in [12], it was neglected. Therefore, we mainly focus on the scheme without Heaviside function in the following.

2.1.2 Mean-Field Limit

Properties of the scheme were investigated on the mean-field level. Up to the author's knowledge, there is no rigorous proof of the limiting equation so far. We therefore have to assume that propagation of chaos holds in order to derive the mean-field equation formally.

Let us assume that the *propagation of chaos* property holds, that means the distribution of all agents X, ν_t^N, satisfies $\nu_t^N \approx \rho_t^{\otimes N}$, $N \gg 1$, and therefore $X^i(t)$ are approximately independently ρ_t-distributed. Then,

$$\frac{1}{N}\sum_{i=1}^{N} \omega_\alpha^f(X^i(t)) \approx \int_{R^d} \omega_\alpha^f(x)d\rho_t, \qquad \frac{1}{N}\sum_{i=1}^{N} X^i(t) w_\alpha^f(X^i(t)) \approx \int_{\mathbb{R}^d} x\omega_\alpha^f(x)d\rho_t,$$

due to the law of large numbers. Hence, $v_f \approx v_f[\rho_t]$ and we obtain the *McKean nonlinear process*

$$d\bar{X}(t) = -\lambda(\bar{X}(t) - v_f[\rho_t])\,dt + \sqrt{2}\sigma|\bar{X}(t) - v_f[\rho_t]|dW_t, \tag{4a}$$

where the weighted average reads

$$v_f[\rho_t] = \frac{1}{\int_{\mathbb{R}^d} \omega_f^\alpha d\rho_t} \int_{\mathbb{R}^d} x\,\omega_f^\alpha d\rho_t, \qquad \rho_t = \mathrm{law}(\bar{X}(t)). \tag{4b}$$

Equation (4a) may be equivalently expressed as the Fokker–Planck equation:

$$\partial_t \rho_t = \Delta(\kappa[\rho_t]\rho_t) + \mathrm{div}(\mu[\rho_t]\rho_t), \tag{5a}$$

$$\kappa[\rho_t](x) = \sigma^2|x - v_f[\rho_t]|^2, \qquad \mu[\rho_t](x) = -\lambda(x - v_f[\rho_t]), \tag{5b}$$

which describes the evolution of the law corresponding to the McKean nonlinear process $\{\bar{X}(t) \in \mathbb{R}^d \mid t \geq 0\}$.

The presence of v_f makes the Fokker–Planck equation nonlinear and nonlocal in both the drift and the diffusion part. This is nonstandard in the literature and raised several analytical and numerical questions that were addressed in [12]. We recall the main results in the following.

2.1.3 Analytical Results for the Original Scheme Without Heaviside Function

The first statement considers the well-posedness of the particle dynamic; see Theorem 2.1 in [12] for the proof.

Theorem 1 *Let the objective function f be locally Lipschitz continuous. For every $N \in \mathbb{N}$, system (3) has a unique strong solution $\{X_t^N : t \geq 0\}$ for any initial condition $X_0^{(N)}$ satisfying $\mathbb{E}|X_0^{(N)}|^2 < \infty$.*

For the original particle scheme, there is neither a proof for consensus formation nor for convergence to the global minimizer. These kinds of results were only addressed on the mean-field level after a formal limiting procedure as $N \to \infty$. A rigorous proof of this limit is open up to the author's knowledge. The following results and some first estimates in the direction of a rigorous proof of the mean-field limit are reported in [12].

Well-posedness of the mean-field equation is established for two classes of objective functions. One result considers only bounded objective functions, and the other result is for objective functions with quadratic growth at infinity. Both versions are based on the following assumption:

Assumption 1 To obtain the well-posedness results of the mean-field equation, we assume that it holds:

1. The cost function $f : \mathbb{R}^d \to \mathbb{R}$ is bounded from below with $\underline{f} := \inf f$.
2. There exist constants L_f and $c_u > 0$ such that

$$\begin{cases} |f(x) - f(y)| \leq L_f(|x| + |y|)|x - y| \quad \text{for all } x, y \in \mathbb{R}^d, \\ f(x) - \underline{f} \leq c_u(1 + |x|^2) \quad \text{for all } x \in \mathbb{R}^d. \end{cases} \tag{A1}$$

Definition 1 We say that a function has *quadratic growth* if there exist constants $M > 0$ and $c_l > 0$ such that

$$f(x) - \underline{f} \geq c_l|x|^2 \quad \text{for } |x| \geq M. \tag{A2}$$

Theorem 2 *Let f be bounded or have quadratic growth, let Assumption 1 hold and $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then there exists a unique nonlinear process $\bar{X} \in C([0, T], \mathbb{R}^d)$, $T > 0$, satisfying*

$$d\bar{X}_t = -\lambda(\bar{X}_t - v_f[\rho_t])\, dt + \sigma|\bar{X}_t - v_f[\rho_t]|dW_t, \qquad \rho_t = law(\bar{X}_t),$$

in the strong sense, and $\rho \in C([0, T], \mathcal{P}_2(\mathbb{R}^d))$ satisfies the corresponding Fokker–Planck equation (5) in the weak sense with $\lim_{t\to 0} \rho_t = \rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$.

Both proofs are based on Schauder's fixed-point argument and can be found in [12]. The main difference of the two versions is the argument for the bound of the second moment. This bound is needed in order to apply Gronwall's theorem and to close the Schauder argument.

Convergence of the scheme towards the global minimizer of the objective function is shown in two steps. The first step assures only the consensus formation. The second one shows that for appropriate parameter choices the consensus location is positioned near the global minimizer. Both results are of asymptotic nature. The consensus formation occurs for $t \to \infty$, and the approximation of the global minimizer depends on the choice of the weight parameter α. For $\alpha \to \infty$, the location of consensus tends towards the global minimizer. For the concentration result, we need this assumption.

Assumption 2 We assume that $f \in C^2(\mathbb{R}^d)$ satisfies additionally

1. $\inf f > 0$.
2. $\|\nabla^2 f\|_\infty \leq c_f$ and there exist constants $c_0, c_1 > 0$, such that

$$\Delta f \leq c_0 + c_1|\nabla f|^2 \qquad \text{in } \mathbb{R}^d.$$

To show the concentration, we investigate the expectation and the variance of the density, which are defined by

$$E(\rho_t) = \int_{\mathcal{X}} x d\rho_t \quad \text{and} \quad V(\rho_t) = \frac{1}{2}\int_{\mathcal{X}} |x - E(\rho_t)|^2 d\rho_t.$$

The details of the concentration procedure are given in [12].

Theorem 3 *Let f satisfy Assumption 2, and let the parameters α, λ and σ satisfy*

$$2\alpha e^{-2\alpha \underline{f}}(c_0\sigma^2 + 2\lambda c_f) < \frac{3}{4}, \qquad 2\lambda b_0^2 - K - 2d\sigma^2 b_0 e^{-\alpha \underline{f}} \geq 0,$$

with $K = V(\rho_0)$ and $b_0 = \|\omega_f^\alpha\|_{L^1(\rho_0)}$. Then $V(\rho_t) \leq V(\rho_0)e^{-qt}$ with

$$q = 2\big(\lambda - (d\sigma^2/b_0)e^{-\alpha \underline{f}}\big) \geq K/b_0^2.$$

Furthermore, there exists a point $\tilde{x} \in \mathbb{R}^d$ for which $E(\rho_t) \to \tilde{x}$ and $v_f[\rho_t] \to \tilde{x}$ as $t \to \infty$.

So far, we just know that the density will concentrate at some point, $\tilde{x}$, and the location of this point remains unknown. Finally, the following result assures that the concentration takes place in a neighbourhood of the global minimizer for appropriately chosen parameters.

Theorem 4 *Let f satisfy Assumption 2. For any given $0 < \epsilon_0 \ll 1$ arbitrarily small, there exist some $\alpha_0 \gg 1$ and appropriate parameters (λ, σ) such that uniform consensus is obtained at a point $\tilde{x} \in B_{\epsilon_0}(x_*)$. More precisely, we have that $\rho_t \to \delta_{\tilde{x}}$ for $t \to \infty$, with $\tilde{x} \in B_{\epsilon_0}(x_*)$.*

Note that due to $W_2(\rho_t, \delta_{\tilde{x}})^2 \leq V(t) + |E(t) - \tilde{x}|^2 \longrightarrow 0$, the convergence of ρ_t towards $\delta_{\tilde{x}}$ is at least in W_2 sense. With this result, we conclude the survey of the analytical results on the original scheme of consensus-based global optimization proposed in [11] and analysed in [12].

2.1.4 Numerical Methods

It is important to notice that the success of the CBO method is far more dependent on the function evaluations than on the accuracy of the numerical scheme. In fact, whenever a particle hits the global minimum of the function, the weighted average v_f is assumed to move to this position and then concentration takes place.

Having this in mind, most of the numerical simulations use basic algorithms such as the Euler–Maruyama scheme [18].

In [11] the formal mean-field limit is underlined by the comparison of numerical results on the particle level with the solution of the candidate equation on the mean-field level. The PDE is solved with the help of a discontinuous Galerkin approach in combination with a Strang splitting. The convective part is solved with the local Lax–Friedrichs scheme and the diffusion part semi-implicitly.

In the following we discuss variants of this method that aim to enhance the performance or extend the class of optimization problems admissible for CBO. We begin with a variant that appears like a slight modification of the above algorithm. However, it has a major impact on the convergence results, especially in high dimensions.

2.2 Variant 1: Component-Wise Diffusion and Random Batches

At first glance, the variant with component-wise independent noise in [13] seems to be a minor modification of the original dynamic. Nevertheless, it turns out that the estimates of the convergence results become independent of the dimension of the state space. This is an advantage, especially when the method is considered for high-dimensional problems, for example, arising in machine learning. In addition to the component-wise diffusion, the authors propose to use mini-batches, a popular approach in many stochastic gradient descent methods [19].

2.2.1 Component-Wise Geometric Brownian Motion

The dynamic with component-wise geometric Brownian motion reads

$$dX^i(t) = -\lambda(X^i(t) - v_f)dt + \hat{\sigma}\sum_{k=1}^{d}(X^i(t) - v_f)_k dW_k^i(t)\mathbf{e}_k, \tag{6}$$

for $i = 1, \ldots, N$ and is supplemented with initial data $X(0) = X_0$. Here $\mathbf{e}_k$ denotes the k-th unit vector in $\mathbb{R}^d$, $(X^i(t) - v_f)_k$ is the k-th entry of the difference and W_k^i are independent standard Brownian motions. The weighted mean, v_f, is given in (2).

Remark 2 Let us mention two differences between (3) and (6):

1. **Component-wise noise:** The component-wise diffusion in (6) scales the distance of X_k^i and v_f element-wise. In case one component of the two coincides, this component of X^i does not change.
2. **Diffusion constants:** The slight difference between the diffusion constants in (3) and (6) $\hat{\sigma} = \sqrt{2}\sigma$ has no significant influence on the performance of the scheme.

The aforementioned dimension independence of the component-wise diffusion can be seen with the help of a simple computation [13]. Let us fix the weighted average v_f at an arbitrary position a. Then, for the dynamics in (3) we find

$$\frac{d}{dt}\mathbb{E}|X(t)-a|^2=-2\lambda\mathbb{E}|X(t)-a|^2+\sigma^2\sum_{i=1}^{d}\mathbb{E}|X(t)-a|^2 = (-2\lambda+\sigma^2 d)\mathbb{E}|X(t)-a|^2.$$

This investigation of the second moment shows that concentration occurs whenever the condition $2\lambda > \sigma^2 d$ is satisfied. In contrast, the same computation for (6) yields

$$\frac{d}{dt}\mathbb{E}|X(t)-a|^2=-2\lambda\mathbb{E}|X(t)-a|^2+\sigma^2\sum_{i=1}^{d}\mathbb{E}|X(t)-a|_i^2 = (-2\lambda+\sigma^2)\mathbb{E}|X(t)-a|^2.$$

The condition for concentration changes to $2\lambda > \sigma^2$. In particular, it is independent of the dimension d.

It can be proven that all estimates needed for the analysis of well-posedness, concentration and convergence towards the global minimizer on the mean-field level are independent of the dimension for the component-wise diffusion variant. Instead of rewriting the statements here, we refer to [13] for all details and proceed with the second interesting feature proposed in the article.

2.2.2 Random Batch Method

The second novelty proposed in [13] is to apply the random-mini batch strategy [20] in two levels: first, instead of evaluating $f(X^i(t))$ for every particle $i = 1, \ldots, N$ in every time step, q random subsets $J^\theta \subset \{1, \ldots, N\}$ with size $|J^\theta| = M \ll N$ and $\theta = 1, \ldots, q$ are drawn and for each of them an empirical expectation $\hat{f}(X^\theta)$ is computed. Based on these function evaluations, a weighted mean is calculated for every batch. Now, one can choose to update the positions of particles by (6) only for the particle in the batch or apply the dynamics to all N particles. For simplicity, we present a version of the algorithm in [13] adapted to the general problem (P). Note that there is an additional parameter, $\gamma_{k,\theta}$, called *learning rate* following the machine learning terminology.

Algorithm 1

Generate $\{X_0^i \in \mathbb{R}^d\}_{i=1}^N$ according to the same distribution ρ_0. Set the remainder set $\mathcal{R}_0$ to be empty. For $k = 0, 1, 2, \ldots$, do the following:

- Concatenate $\mathcal{R}_k$ and a random permutation $\mathcal{P}_k$ of the indices $\{1, 2, \ldots, N\}$ to form a list $\mathcal{I}_k = [\mathcal{R}_k, \mathcal{P}_k]$. Pick $q = \lfloor \frac{N+|\mathcal{R}_k|}{M} \rfloor$ sets of size $M \ll N$ from the list $\mathcal{I}_k$ in order to get batches $B_1^k, B_2^k, \ldots, B_q^k$ and set the remaining indices to be $\mathcal{R}_{k+1}$. Here, $|\mathcal{R}_k|$ means the number of elements in $\mathcal{R}_k$.
- For each B_θ^k ($\theta = 1, \ldots, q$), do the following:

  1. Calculate the function values (or approximated values) of f at the location of the particles in B_θ^k by $f^j := f(X^j), \ \forall j \in B_\theta^k$.
  2. Update $v_{k,\theta}$ according to the following weighted average:

$$v_{k,\theta} = \frac{1}{\sum_{j \in B_\theta^k} \mu_j} \sum_{j \in B_\theta^k} X^i \mu_j, \quad \text{with} \quad \mu_j = e^{-\alpha f^j}.$$

3. Update X^j for $j \in \mathcal{J}_{k,\theta}$ as follows:

$$X^j \leftarrow X^j - \lambda\gamma_{k,\theta}(X^j - v_{k,\theta}) + \sigma_{k,\theta}\sqrt{\gamma_{k,\theta}} \sum_{i=1}^d \mathbf{e}_i (X^j - v_{k,\theta})_i \, z_i^j, \;\; z_i^j \sim \mathcal{N}(0,1),$$

where $\gamma_{k,\theta}$ is chosen suitably and there are two options for $\mathcal{J}_{k,\theta}$:

$$\textit{partial updates}: \quad \mathcal{J}_{k,\theta} = B_\theta^k, \quad \text{or} \quad \textit{full updates}: \quad \mathcal{J}_{k,\theta} = \{1, \ldots, N\}.$$

- Check the **Stopping criterion:**

$$\frac{1}{d} \|\Delta x\|_2^2 \leq \epsilon,$$

where $\|\cdot\|_2$ is the Euclidean norm and Δv is the difference between two most recent $v_{k,\theta}$. If this is not satisfied, repeat.

Note that due to the mini-batch evaluation additional noise is added to the algorithm. The authors discuss in [13] that this additional noise causes the algorithm to work fine even without the geometric Brownian motion. For details and additional ideas on how to improve the convergence for objective functions with a typical machine learning structure, we refer to [13].

We conclude this section with some ideas on the numerical implementation and the performance of the algorithm with random batches and component-wise geometric Brownian motion.

2.2.3 Implementation and Numerical Results

A typical challenge is to avoid overshooting, which refers to oscillations around v in our context. The authors propose two approaches to do so.

First, the drift and diffusion parts of the scheme can be split. Then the drift part can be computed explicitly using

$$\hat{X}_k^j = v_k + (X_k^j - v_k)e^{-\lambda\gamma},$$

which corresponds to a scheme for solving the ODE $dX^j = -\lambda(X^j - v)$ on the interval $t \in [k\gamma, k(\gamma + 1)]$. The diffusion update is given by

$$X_{k+1}^j = \hat{X}_k^j + \sigma\sqrt{\gamma}\sum_{i=1}^{d} \mathbf{e}_i \left(\hat{X}_k^j - v\right)_i z_i^j .$$

Second, they propose to freeze the weighted average over fixed time intervals. On each of these intervals, the geometric Brownian motion can be solved by

$$X_{k+1}^j = v + \sum_{i=1}^{d} \mathbf{e}_i \left(\hat{X}_k^j - v\right)_i \exp\left(\left(-\lambda - \frac{1}{2}\sigma^2\right)\gamma + \sigma\sqrt{\gamma} z_i^j\right).$$

Moreover, they report that the splitting and the freezing approach lead to comparable results in most numerical simulations. For more details, see [13].

The aforementioned paper reports results of three numerical studies. The first is a proof of concept using a one-dimensional objective function with many local minima and oscillatory behaviour. The second study compares the method to the performance of a stochastic gradient descent method applied to the MNIST data set. Finally, results for a test function in high dimensions with many local minima are provided.

The test cases show that the proposed CBO algorithm with component-wise Brownian motion and mini-batches outperforms the stochastic gradient descent algorithm. Moreover, it turns out that the version with mini-batches leads to better results than the one with full evaluations in case of the MNIST data set. For more detailed discussions and studies of the influence of α and N on the performance, we refer the reader to the original article [13].

2.3 Variant 2: Component-Wise Common Diffusion

The idea of component-wise diffusion plays a role as well in [14, 21] with the main difference that the component-wise noise is *common* for all particles, that means the dynamic is given by

$$dX^i(t) = -\lambda(X^i(t) - v_f)dt + \hat{\sigma}\sum_{k=1}^{d}(X^i(t) - v_f)_k dW_k(t)\mathbf{e}_k, \tag{7}$$

where W_k are i.i.d. one-dimensional Brownian motions. The dynamic is supplemented with initial conditions $X^i(0) = X_0^i$ and v_f as above. Note that the Brownian motion does not depend on the specific particle i and therefore all particles encounter a common noise.

In addition to the continuous-time particle scheme given above, the article discusses a time-discrete version. Let $h > 0$ denote the time step, i.e. $t = nh$, we set $X_n^i := X^i(nh)$. The discrete algorithm reads

$$X_{n+1}^i = X_n^i - \lambda h (X_n^i - v_f) + \sigma\sqrt{h}\sum_{k=1}^{d}(X_n^i - v_f)_k Z_n^k \mathbf{e}_k, \tag{8}$$

where $\{Z_n^k\}_{n,k}$ are i.i.d. standard normal distributed random variables, $Z_n^k \sim \mathcal{N}(0, 1)$. Note that compared to [14] the notation was adjusted for the sake of a consistent presentation.

2.3.1 Analytical Results

The common noise approach has the advantage that a convergence study can be done directly on the level of particles without passing to the mean-field level. Similar to the strategy of the proof on the mean-field level, the convergence proof for the common noise scheme is split into two parts: first, under certain conditions on the drift and diffusion parameters, a general convergence to consensus result for $t \to \infty$ is shown. In a second step the authors provide sufficient conditions on the system parameters and initial data, which guarantee that the location of the consensus is in a small neighbourhood of the global minimum almost surely. The conditions on the parameters are independent of the dimension similar to Variant 1 (see Sect. 2.2).

Despite these two main results, some properties of the continuous and discrete deterministic schemes are discussed. In fact, it is proven that the convex hull of the particles following the deterministic (both time-continuous and time-discrete) schemes are contractive as time evolves. The convergence to a consensus state is a direct consequence.

The same contraction property is not given for the scheme with noise. Nevertheless, for the common noise approach the relative difference of two particles satisfies a geometric Brownian motion. Hence, an exact solution can be established using stochastic calculus. This implies that the relative state difference converges almost surely. The details of the theorem are as follows.

Theorem 5 *Let $X^i(t)$ be the i-th agent of a solution to (7). Then for $i \neq j = 1, \ldots, N$ it holds*

$$\mathbb{E}|X^i(t) - X^j(t)|^2 = e^{-(2\lambda-\sigma^2)t}\mathbb{E}|X_0^i - X_0^j|^2, \quad t > 0.$$

In particular L^2-consensus emerges if and only if $\lambda - \frac{\sigma^2}{2} > 0$.

A similar result is obtained for the time-discrete dynamic (8). The condition for the convergence depends on σ, λ and h. In fact, several different conditions are discussed. For details, we refer to [14].

The second step that shows that for well-chosen parameters the consensus state is located in a neighbourhood of the global minimizer is more involved. Here, we only state the main result that needs the following assumption.

Assumption 3 We assume that f and the initial conditions satisfy:

1. $f \in C_b^2(\mathbb{R}^d)$ with $\inf_{x\in\mathbb{R}^d} f(x) > 0$ and

$$C_L := \max\left\{ \sup_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\|_2, \max_{1\le l\le d} \sup_{x\in\mathbb{R}^d} |\partial_l^2 f(x)| \right\} < \infty.$$

2. For some $\epsilon \in (0, 1)$, the initial conditions X_0^i are i.i.d. with $X_0^i \sim X_{\text{in}}$ for some random variable X_{in} that satisfies

$$(1-\epsilon)\mathbb{E}[e^{-\alpha f(X_{\text{in}})}] \ge \frac{2\lambda+\sigma^2}{2\lambda-\sigma^2} C_L \alpha e^{-\alpha f(X_*)} \sum_{l=1}^{d} \mathbb{E}\left[\max_{1\le i\le N} (X_0^i - v_f(0))_l \right].$$

Theorem 6 *Let Assumption 3 hold and suppose* $2\lambda > \sigma^2$*. Then for a solution* X *to* (7) *it holds*

$$\lim_{t\to\infty} \operatorname{essinf}_\omega f(X_t^i(\omega)) \le \operatorname{essinf}_\omega f(X_{in}(\omega)) + E(\alpha)$$

for some function $E(\alpha)$ *with* $\lim_{\alpha\to\infty} E(\alpha) = 0$*. In particular, if the global minimizer* X_* *is contained in the support of law*$(X_{in}) = \rho_0$*, then*

$$\lim_{t\to\infty} \operatorname{essinf}_\omega f(X_t^i(\omega)) \le f(X_*) + E(\alpha).$$

The convergence of the time-discrete algorithm was not established in [14] due to the lack of a discrete analogue of Itô's stochastic calculus. In a subsequent article [21] the authors give an elementary convergence and error analysis for the time-discrete version (8) under some additional regularity conditions on f. Moreover, exponential decay rates of the distances between the particles are established. The proofs are rather technical and go beyond the scope of this survey. We therefore refer the interested reader to [21].

2.3.2 Numerical Results

A priori it is not clear how the common noise algorithm performs compared to the well-tested component-wise noise version in Sect. 2.2. In [14] some numerical results of the common noise algorithm are provided. They underline the analytical results on the convergence of the distance of two particles and indicate that also the common noise version leads to reasonable results. A large-scale comparison of Variant 1 and the common noise scheme of this section is missing up to the author's knowledge.

3 Relationship of CBO and Particle Swarm Optimization

Consensus-based optimization is inspired by Particle Swarm Optimization (PSO) schemes [9]. It is worthwhile to compare the methods to gain further insight to their behaviour, performance and the qualities. Let us recall the formulation of the PSO dynamic [22]: the update for the i-th particle is given by

$$X^i \leftarrow X^i + V^i, \quad i = 1, \dots, N$$

$$V^i \leftarrow \omega V^i + \sum_{k=1}^{d} \Big(U^i_{1,k} (p_{\text{personal}} - X^i)_k + U^i_{2,k} (p_{\text{global}} - X^i)_k \Big),$$

where $U^i_{1/2}$ are d-dimensional vectors of random numbers, which are uniformly distributed in $[0, \phi_1]$ and $[0, \phi_2]$, respectively. p_{global} denotes the best position that any of the particles has seen, and p_{personal} denotes the best position particle i has seen. The parameters $\phi_{1,2}$ define the magnitude of the stochastic influences, and ω can be interpreted as inertia parameter. V^i is originally kept within box constraints, given by the range $[-V_{\max}, V_{\max}]$. In contrast to the first-order dynamic of CBO, PSO is of second order, which may lead to inertia effects. Moreover, the stochastic influence does not vanish, and therefore one cannot expect any kind of consensus formation. The approximation of the global best is p_{global} whenever the PSO algorithm is stopped. The global best information in PSO prevents a direct passage to the mean-field limit.

The main ingredient of CBO is the weighted average v_f. For $\alpha \gg 1$, it can be interpreted as an approximation of the current best particle position. Here, we use best in the sense that the function value is the lowest compared to the function values of all other particles. This current best particle does move only slightly, as its distance to v_f is small and therefore the drift and diffusion terms are small. Therefore the current best particle can as well be interpreted as the global best position seen so far. Hence, v_f is the analogue of p_{global} in PSO. In addition, the PSO dynamic includes the so-called local best position, which refers to the best position that each of the particles has seen. This local best is modelled in [15] using a memory effect. The same local best is mentioned as well in a recent preprint [23], which additionally considers a continuous description of PSO and computes the corresponding macroscopic equations to clarify the relationship of PSO and CBO. The details of the CBO with local or personal best information are given in the following.

3.1 Variant 4: Personal Best Information

Consensus-based optimization with global and local best in the sense of PSO is proposed in [15] and based on the component-wise diffusion variant (see Sect. 2.2).

The dynamic reads as follows:

$$dX^i(t) = \Big[-\lambda(t,X)(X^i(t)-v_f) - \mu(t,X)(X^i(t)-p^i(t))\Big]\,dt \\ + \sqrt{2}\sigma\sum_{k=1}^{d}(X^i(t)-v_f)_k dW_k^i(t)\mathbf{e}_k, \qquad i=1,\dots,N, \tag{9}$$

with v_f as given above and the personal best is modelled by

$$p^i(t) = \begin{cases} X_0^i, & t=0,\\ \int_0^t X^i(s)\exp(-\beta f(X^i(s)))ds \Big/ \int_0^t \exp(-\beta f(X^i(s)))ds, & \text{otherwise.}\end{cases}$$

This personal best approximation uses the same idea as v_f but with respect to time in contrast to the integral over the state space. Again by Laplace principle (see Proposition 1), we expect that $p^i(t)$ approximates the best position that particle X^i has seen up to time t.

Remark 3 To circumvent the integral over time, it is tempting to rewrite the numerator and denominator of p^i as SDE. Notice that the initial condition of each personal best would need to be positioned at zero in order to obtain the exact definition above.

To make sure that particles do not get stuck in the middle, each particle has to choose whether it moves towards v_f or towards its personal best p^i. As we aim for a global minimizer, we assume that this decision is based on the cost functional values, which motivates to set the prefactors λ and μ as

$$\lambda(t,X) = \mathrm{H}(f(X^i(t)) - f(v_f))\,\mathrm{H}(f(p^i) - f(v_f)),$$
$$\mu(t,X) = \mathrm{H}(f(X^i(t)) - f(p^i))\,\mathrm{H}(f(v_f) - f(p^i)).$$

This is leading to the following behaviour:

- If $f(v_f)$ is smaller than $f(X^i)$ and $f(p^i)$, the particle moves towards v_f.
- If $f(p^i)$ is smaller than $f(X^i)$ and $f(v_f)$, the particle moves towards p^i.
- If none of the above holds, the particle still explores the function landscape via Brownian motion until it reaches the global best v_f.

Using a regularized version of the Heaviside function, H^ϵ, the well-posedness of the above system is proven in [15]. There are no mean-field result and no convergence result reported.

3.1.1 Performance

Note that the additional evaluation of the personal best position has minor impact on the computational costs, as the time integrals in p^i allow for an accumulative computation. Note further that even though the Heaviside function needs to be regularized for the analysis, the numerical results can work with the original Heaviside formulation.

The numerical results indicate that the personal best information raises the probability of finding the global best position, if few particles are involved in the search. As the number of particles needed for satisfying results depends on the dimension of the state space, this result is particularly important in high dimensions. If the number of particles is large enough, no significant influence of the personal best information is noted.

4 CBO with State Constraints

Many global optimization tasks have a constrained state space. The simplest version of constraints are box constraints. These can be included into each of the aforementioned CBO versions by projecting particles back into the box, whenever they are about to leave it.

The situation is more complicated, when the state space is given in the form of a hyper-surface of $\mathbb{R}^d$. For example the sphere

$$\mathbb{S}^2 = \{x \in \mathbb{R}^3 \colon |x| = 1\}$$

is a hyper-surface of $\mathbb{R}^3$. In [16, 24] a variant of CBO on such hyper-surfaces is proposed. The first paper is concerned with the well-posedness and the mean-field limit of the variant, and the second article discusses the convergence to global minimizers and applications in machine learning. A major advantage of this variant is the fact that compactness is assured by the constraint. Therefore the mean-field limit can be established rigorously. In the following we discuss the details [16].

4.1 Variant 5: Dynamics Constrained to Hyper-Surfaces

The restriction to the hyper-surface leads to a new formulation of the optimization problem

$$\min_{x\in\Gamma} f(x),$$

where Γ represents the hyper-surface and $f: \mathbb{R}^d \to \mathbb{R}$ as above. We assume that Γ is a connected and smooth compact hyper-surface embedded in $\mathbb{R}^d$, which is represented as zero-level set of a signed distance function γ with $|\gamma(x)| = \operatorname{dist}(x, \Gamma)$ leading to

$$\Gamma = \{x \in \mathbb{R}^d : \gamma(x) = 0\}.$$

If $\partial\Gamma = \emptyset$, we assume for simplicity that $\gamma < 0$ on the interior of Γ and $\gamma > 0$ on the exterior. The gradient, $\nabla\gamma$, is the outward unit normal on Γ whenever γ is defined. In addition, we assume that there exists an open neighbourhood $\hat{\Gamma}$ of Γ such that $\gamma \in C^3(\hat{\Gamma})$. All these assumptions allow us to work with the Laplace–Beltrami operator. For example, for the sphere $\mathbb{S}^{d-1}$ we can choose

$$\gamma(x) = |x| - 1 \quad \text{with} \quad \nabla\gamma(x) = \frac{x}{|x|} \quad \text{and} \quad \Delta\gamma(x) = \frac{d-1}{|x|}.$$

In [16] a Kuramoto–Vicsek-type dynamic is proposed as

$$\begin{aligned} dX^i(t) &= -\lambda P(X^i(t))(X^i(t) - v_f)dt + \sigma|X^i(t) - v_f|P(X^i(t))dB^i(t) \\ &\quad - \frac{\sigma^2}{2}(X^i(t) - v_f)^2 \Delta\gamma(X^i(t))\nabla\gamma(X^i(t))dt, \quad i = 1, \dots, N, \end{aligned} \tag{10}$$

with initial condition $X(0) = X_0$. In contrast to the aforementioned schemes there appears a projection operator P defined by

$$P(x) = I - \nabla\gamma(x)\nabla\gamma(x)^T.$$

For the sphere, we obtain the $P(x) = I - \frac{xx^T}{|x|^2}$. In addition to this projection there appears a third term in (10). The two mechanisms ensure that the dynamics stays on the hyper-surface Γ.

Remark 4 Note that the dynamic is described in $\mathbb{R}^d$. On the one hand this allows for a simple statement of the scheme. On the other hand it is likely that the weighted average is not positioned at Γ, i.e. $v_f \notin \Gamma$. This is caused by the averaging of particles on a hyper-surface. Nevertheless, for $\alpha \gg 1$, v_f approximates the current best particle, which is contained in Γ due to the projection and correction terms.

The constraint enables us to give rigorous arguments for the limit $N \to \infty$, which results in the nonlocal, nonlinear Fokker–Planck equation:

$$\partial_t \rho_t = \lambda \nabla_\Gamma \cdot [P(x)(x - v_f)\rho_t] + \frac{\sigma^2}{2}\Delta_\Gamma(|x - v_f|^2 \rho_t), \quad t > 0, x \in \Gamma,$$

with initial condition $\rho_0 \in \mathcal{P}(\Gamma)$. The operators ∇_Γ and Δ_Γ denote the divergence and Laplace–Beltrami operator on the hyper-surface Γ, respectively. In the following we summarize the analytical results for this variant, which are reported in [16].

4.1.1 Analytical Results

The following analytical results focus on the well-posedness and the rigorous mean-field limit of the constrained scheme.

As the dynamic is living in $\mathbb{R}^d$, there are some technical issues with P, $\Delta\gamma$ and $\nabla\gamma$. In fact, these are not defined for $x = 0$ and the authors propose to replace them with regularizations. Moreover, a regularized extension of f, called $\tilde{f}$, is introduced.

Assumption 4 Let $\tilde{f}$ be globally Lipschitz continuous and such that it holds:

$$\tilde{f}(x) = f(x) \text{ for } x \in \hat{\Gamma},$$

$$\tilde{f}(x) - \tilde{f}(y) \leq L|x-y| \quad \text{for all } \; x, y \in \mathbb{R}^d \text{ for } L > 0,$$

$$-\infty < \underline{\tilde{f}} := \inf \tilde{f} \leq \tilde{f} \leq \sup \tilde{f} := \overline{\tilde{f}} < +\infty.$$

The authors emphasize that the regularization $\tilde{f}$ is introduced only for technical reasons and that it does not the influence the optimization problem, as it can be shown that the dynamic stays on the hyper-surface whenever it is initialized there.

The well-posedness results for the particle and the mean-field scheme [16] read as follows.

Theorem 7 *Let Assumption 4 hold and f with $0 \leq f$ be locally Lipschitz. Moreover, let $\rho_0 \in \mathcal{P}(\Gamma)$. For every $N \in \mathbb{N}$, there exists a path-wise unique strong solution $X = (X^1, \dots, X^N)$ to the system (10) with initial condition $X(0) = X_0$. Moreover, it holds $X^i(t) \in \Gamma$ for all $i \in \{1, \dots, N\}$ and $t > 0$.*

The well-posedness of the PDE is established similar to Theorem 2 in Sect. 2.2 with the help of an auxiliary mono-particle process $\bar{X}$ satisfying

$$\begin{aligned} d\bar{X}(t) = -\lambda P(\bar{X}(t))(X(t) - v_f)dt + \sigma|\bar{X}(t) - v_f|P(\bar{X}(t))dW(t) \\ - \frac{\sigma^2}{2}(\bar{X}(t) - v_f)^2 \Delta\gamma(\bar{X}(t))\nabla\gamma(\bar{X}(t))dt, \end{aligned} \tag{11}$$

in strong sense for any initial data $\bar{X}(0) \in \Gamma$ distributed according to $\rho_0 \in \mathcal{P}(\Gamma)$. It holds $\text{law}(\bar{X}(t)) = \rho_t$, which allows to define $v_f = v_f[\rho]$ as in (4b). For details on the well-posedness of the PDE, we refer the interested reader to [16]. We proceed with the ideas leading to the rigorous mean-field limit result.

Using N independent copies of this mono-particle process allows to obtain the rigorous mean-field limit with the well-known technique of Sznitman [25]. In contrast to the unconstrained case, where the rigorous proof of the mean-field limit is open, the compactness of the hyper-surface makes the difference.

Theorem 8 *Let Assumption 4 hold and f be locally Lipschitz. For any $T > 0$, let $X^i(t)$ and $\bar{X}^i(t)$, $i = 1, \ldots, N$, be solutions to (10) or the corresponding mono-particle process, respectively, up to time T with the same initial data $X^i(0) = \bar{X}^i(0)$ and the same Brownian motions $W^i(t)$. Then there exists a constant $C > 0$ depending only on parameters, regularizations and constants, such that*

$$\sup_{i=1,\ldots,N} \mathbb{E}[|X^i(t) - \bar{X}^i(t)|^2] \leq \frac{CT}{N}\Big(1 + CTe^{CT}\Big)$$

holds for all $0 \leq t \leq T$.

Remark 5 In addition to these results, there is a preprint [24] that reports on the convergence to the global minimizer and simulation results for applications in machine learning for the CBO scheme constrained to hyper-surfaces (10).

With this, we conclude the survey of the variants. In the next section we briefly discuss some applications and performance results of the variants.

5 Overview of Applications

The CBO variants were studied in various test problems. Initially, benchmark functions from global optimization were used to get first results. As the variants are tailored for high-dimensional applications arising in machine learning problems, they are tested against stochastic gradient descent. The preprint [24] shows some first results for the constraint method for the Ackley function on the sphere and machine learning scenarios.

5.1 Global Optimization Problems: Comparison to Heuristic Methods

In [11, 26] benchmark functions from global optimization with various local minima and only one global minimum such as the Ackley, Rastrigin, Griewank, Zakharov and Wavy function were employed to test CBO against PSO and WDO. It turns out that CBO shows the best overall performance. In particular in scenarios where PSO and WDO have a very low success rate, CBO leads to reasonable results with success rates $>50\%$.

5.2 Machine Learning

Variants 2 and 3 were tailored for applications in machine learning. A comparison between Variant 2 and the stochastic gradient descent is reported in [13]. For a global optimization problem with an objective function that has many local minima, the CBO variant outperforms SGD in terms of the success rate. The authors explain that this is caused by the fact that SGD needs a lot of time to escape from basins of local minima. A comparison with fixed computational cost is missing up to author's knowledge.

Another test case considers the well-known MNIST data set. Here, the differences between Variant 2 and SGD are less obvious. Nevertheless, CBO leads to slightly better results. See [13] for more details.

5.3 Global Optimization with Constrained State Space

The preprint [24] investigates global optimization problems from signal processing and machine learning that are naturally stated on the sphere. The first one is *phase retrieval*, where the task is to recover an input vector z from noisy quadratic measurements. The simulation results show that Variant 5 is able to match state-of-the-art methods for phase retrieval.

The second applications is *robust subspace detection*. Here, the task is to find the principal component of a given point cloud. The performance of Variant 5 is reported to be equally good as the one of the Fast Median Subspace method applied to synthetic data. Then a computation of eigenfaces based on real-life photos from the *10k US Adult Faces Database* is studied. It turns out that the results of Variant 5 are more reliable than the ones by SVD when outliers are present in the data set.

5.4 PDE Versus SDE Simulations

In many applications of statistical physics, for example, particles in a plasma, electrostatic force or vortices in an incompressible fluid in two space dimensions, the mean-field equation is used to reduce computational cost [27]. For consensus-based optimization, we strongly recommend using the particle level for simulations. This is due to the fact that not too many particles are needed for reasonable results and for high-dimensional problems the computation of the PDE solution is infeasible. The comparison of SDE and PDE results shown in [11] is just to underline the formal limit numerically and thus to justify the analysis of the scheme on the PDE level.

6 Conclusion, Outlook and Open problems

In this survey we collected the main results on consensus-based optimization algorithms. First, we stated the original scheme on the particle level and the analytical results after a formal mean-field limit. Then we discussed variants with component-wise independent and common noise and mini-batch approaches that are tailored for high-dimensional applications arising from machine learning. A variant with component-wise common noise allows for analytical results on the particle level without passing to the limit $N \to \infty$.

Consensus-based optimization has similarities to the well-known Particle Swarm Optimization algorithms. Those were addressed in Sect. 3, where we considered a variant that involves the personal best state of each particle in the dynamic. The survey on the variants was completed with a section on the variants for constrained global optimization problems, which involves the divergence and Laplace–Beltrami operator for hyper-planes. Then we shortly summarized some performance results of the CBO variants and mentioned comparisons to PSO, WDO and SGD. We conclude the survey with remarks on recent preprints and open problems.

Recent Preprints

This survey chapter discusses recent advances of the CBO model that have been published in peer-reviewed journals. Despite these, there are some preprints available, which have not passed the peer review at the time of the final version of this survey:

- A recent preprint [28] proposes CBO with adaptive momentum estimation (ADAM) scheme, which is well-known in the community of stochastic gradient descent methods. The article claims that the new scheme has high success rates at a low cost. Moreover, it can handle nondifferentiable activation functions in neural networks.
- As mentioned in Sect. 3, there is a preprint [23] that discusses a SDE version of the PSO model that allows for passing to the limit $N \to \infty$. A formal analysis on the mean-field level compares properties of CBO and PSO.
- In Sect. 4.1 the preprint concerning the convergence to the global minimizer and machine learning application for CBO constrained to hyper-surfaces [24] was mentioned.
- Preprint [29] discusses a rigorous proof of the mean-field limit. Compactness is established using probabilistic and stochastic arguments.
- An alternative proof of the convergence to the global optimizer is provided in [30]. The results show that CBO performs a convexification of a very large class of optimization problems, as the number of optimizing agents goes to infinity. Moreover, the article proves a quantitative nonasymptotic Laplace principle, which may be of independent interest.
- A sampling approach based on CBO ideas is proposed in [31]. The method allows for the generation of approximate samples from a given target distribution.

Interesting applications are the determination of the maximum a posteriori estimator and sampling from the Bayesian posterior.

Open Problems
Let us mention some interesting open problems in the context of CBO:

– Despite the fact that a rigorous mean-field limit for the unconstrained method in $\mathbb{R}^d$ was established in [29], a quantitative convergence rate is still open.
– A convergence analysis on the particle level was only done for the componentwise common noise algorithm. A rigorous convergence analysis for other variants on the particle level remains open.
– For comparison and qualitative performance results, an estimate on the speed of convergence of the particles to the consensus-point would be of great interest. This point was mentioned in [21] and is still open up to the author's knowledge.
– In most application the structure of the objective functions is unknown, and therefore one cannot guarantee the existence of a unique global minimum. This can lead to difficulties with v_f for symmetric objective functions. A symmetry breaking generalization to problems with multiple global minima would therefore be very interesting.

Altogether, the analytical results and the numerical performance of the CBO variants are very promising and motivate for further research.

References

1. Mohan, B. C. and Baskaran, R.: A survey: Ant colony optimization-based recent research and implementation on several engineering domain. Expert Syst. Appl. **39**, 4618–4627 (2012)
2. Karaboga, D., Gorkemli, B., Ozturk, C., and Karaboga, N.: A comprehensive survey: Artificial bee colony (ABC) algorithm and applications. Artif. Intell. Rev. **42**, 21–57 (2014)
3. Yang, X.-S.: Firefly Algorithms for Multimodal Optimization. In: Stochastic Algorithms: Foundations and Applications, SAGA 2009, Lecture Notes in Computer Sciences, Vol. 5792, pp. 169–178 (2009)
4. Jamil, M. and Yang, X.-S.: A literature survey of benchmark functions for global optimisation problems. Int. J. Math. Model. Numer. Optim. **4**, 150–194 (2013)
5. Bayraktar, Z., Komurcu, M. Bossard, J.A. and Werner, D.H.: The Wind Driven Optimization Technique and its Application in Electromagnetics. In: IEEE Transactions on Antennas and Propagation **61**, 2745–2757 (2013)
6. Henderson, D., Jacobson, S.H. and Johnson, A.W.: The theory and practice of simulated annealing. In: Handbook of Metaheuristics, International Series in Operations Research & Management Science, **57** 287–319, Springer, Boston (2003)
7. Monmarché, P.: Hypocoercivity in metastable settings and kinetic simulated annealing, Probab. Theory Relat. Fields **172**,1215–1248 (2018)
8. Chak, M., Kantas, N. and Pavliotis, G.A.: On the Generalised Langevin Equation for Simulated Annealing, arXiv:2003.06448v3 (2021)
9. Kennedy, J. and Eberhart, R.C.: Particle Swarm Optimization. Proc. IEEE Int. Conf. Neu. Net. **4**, 1942–1948 (1995)
10. Hegselmann, R. and Krause, U.: Opinion dynamics and bounded confidence models, analysis, and simulation. J. Artif. Soc. Social Simulat. **5**, 1–33 (2002)

11. Pinnau, R., Totzeck, C., Tse, O. and Martin, S.: A consensus-based model for global optimization and its mean-field limit. Math. Meth. Mod. Appl. Sci. **27**,183–204 (2017)
12. Carrillo, J.A., Choi, Y.-P., Totzeck, C. and Tse, O.: An analytical framework for a consensus-based global optimization method. Math. Mod. Meth. Appl. Sci. **28**, 1037–1066 (2018)
13. Carrillo, J.A., Jin, S., Li, L. and Zhu, Y.: A consensus-based global optimization method for high dimensional machine learning problems. ESAIM: COCV **27**, S5 (2021)
14. Ha, S.-Y. and Jin, S. and Kim, D.: Convergence of a first-order consensus-based global optimization algorithm. Math. Mod. Meth. Appl. Sci. **30**, 2417–2444 (2020)
15. Totzeck, C and Wolfram, M.-T.: Consensus-Based Global Optimization with Personal Best. Math. Biosci. Eng. **17**, 6026–6044 (2020)
16. Fornasier, M. and Huang, H. and Pareschi, L. and Sünnen, P.: Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. Math. Mod. Meth. Appl. Sci. **30**, 2725–2751 (2020)
17. Dembo, A. and Zeitouni, O.: Large Deviations Techniques and Applications, Applications of Mathematics Vol. 38, Springer Science and Business Media (2009)
18. Higham, D.J.: An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations. SIAM Rev. **43**, 525–546 (2001)
19. Robbins, H. and Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
20. Jin, S. and Li, L., Liu, J.-G.: Random batch methods (RBM) for interacting particle systems. J. Comput. Phys. **400**, 108877 (2020)
21. Ha, S.-Y. and Jin, S. and Kim, D.: Convergence and error estimates for time-discrete consensus-based optimization algorithms. Numer. Math. **147**, 255–282 (2021)
22. Poli, R. and Kennedy, J. and Blackwell, T.: Particle swarm optimization - An overview. Swarm Intell. **1**, 33–57 (2007)
23. Grassi, S. and Pareschi, L.: From particle swarm optimization to consensus based optimization: stochastic modeling and mean-field limit. arXiv:2012.05613 (2020)
24. Fornasier, M. and Huang, H. and Pareschi, L. and Sünnen, P.: Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. arXiv:2001.11988 (2020)
25. Sznitman, A.-S.: Topics in propagation of chaos. In: Ecole d'été de probabilités de Saint-Flour XIX – 2089, 165–251. Springer (1991)
26. Totzeck, C., Pinnau, R., Blauth, S. and Schotthöfer, S.: A Numerical Comparison of Consensus-Based Global Optimization to other Particle-based Global Optimization Schemes. PAMM **18**, e201800291 (2018)
27. 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, pp. 1–144. Springer (2016)
28. Chen, J., Jin, S. and Lyu, L.: A Consensus-based global optimization method with adaptive momentum estimation, arXiv:2012.04827 (2020)
29. Huang, H. and Qiu, J.: On the mean-field limit for the consensus-based optimization. arXiv:2105.12919v1 (2021)
30. Fornasier, M., Klock, T. and Riedl, K.: Consensus-based optimization methods converge globally in mean-field law, arXiv:2103.15130v3 (2021)
31. Carrillo, J.A., Hoffmann, F., Stuart, A.M. and Vaes, U.: Consensus Based Sampling. arXiv:2106.02519v1 (2021)