

ABSTRACT

Title of Dissertation: **SWARM DYNAMICS: FLOCKING AND
COMMUNICATION-BASED OPTIMIZATION**

Jingcheng Lu
Doctor of Philosophy, 2023

Dissertation Directed by: **Professor Eitan Tadmor**
Department of Mathematics

In this thesis, we discuss the swarm dynamics in the modeling of flocking behavior and in the application to numerical optimization. In general, the research of swarm dynamics deals with the collective behavior of multi-agent systems, where interactions between individuals would lead to the emergence of higher-level, self-organized structures of large crowd. Examples occur in a variety of research fields, including physics, biology, human society, and artificial intelligence, etc. Due to the wide range of applications, swarm dynamics has been an important research topic over the last decades.

The first part of our study is devoted to the flocking dynamics driven by the Cucker-Smale system. The flocking behavior is addressed at particle and hydrodynamic levels for single- and multi-species. When studying the hydrodynamic model, a canonical issue is the lack of closure for pressure. The main contribution of our work is to prove the hydrodynamic alignment for multi-species without assuming specific thermodynamic closure. The proof is based on the graph connectivity and a weak dispersion bound. Besides the analytical results, we also discuss the

simulation techniques for collective dynamics, including particle and hydrodynamic simulations. Numerical examples are presented to demonstrate the flocking behavior.

The second part of the thesis focuses on the swarm-based approach for numerical optimization. Based on the idea of consensus formation, we propose a new Swarm-Based Gradient Descent (SBGD) method. To improve the flexibility of exploration, the method employs multiple interacting agents to search for solution. The key innovation of SBGD is to introduce individual masses, which distinguish the quality of different agents and dictate their choice of step size. Agents exchange information through dynamical mass transitions. As iterations proceed, the good agents will gradually become heavy ‘leaders’ and help the crowd find better minimizers. The remarkable improvement of SBGD over the classical gradient descent method and the recently popular Adam method is verified in a series of benchmark tests.

SWARM DYNAMICS: FLOCKING AND COMMUNICATION-BASED OPTIMIZATION

by

Jingcheng Lu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:

Professor Eitan Tadmor, Chair/Advisor
Professor James D. Baeder, Dean's Representative
Professor Trivisa Konstantina
Professor Maria K. Cameron
Dr. Anil Colpan Zenginoglu

© Copyright by
Jingcheng Lu
2023

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost, I'd like to thank my advisor, Professor Eitan Tadmor, for his persistent support in my doctoral study. This thesis would not have been possible without his tremendous help. Throughout my journey as a Ph.D. student, he has provided me with many invaluable insights on research topics, including the inspiring idea of the new swarm-based approach for numerical optimization, as well as the new angle of analyzing collective hydrodynamics. I also appreciate his patient help on improving my academic writing and presentation skills. Besides mathematical research, he also gave me many valuable suggestions in life and corrected my immature behaviors. Having Professor Tadmor as my advisor is one of the luckiest things in my life.

I would like to thank Professor James Baeder for his guidance in my study of computational fluid dynamics. His CFD courses helped me build up a good physical insight on the numerical methods for PDEs. I also appreciate his valuable guidance in my thesis research for dual Master's degree in aerospace engineering, which is another important milestone of mine. Interaction with Professor Baeder is always a pleasure.

I would like to thank Dr. Anil Zenginoglu for his collaboration on the research project for swarm-based gradient descent method. His precious experience and suggestions on the compu-

tational aspect are highly appreciated. I am grateful for his help in verifying the performance of SBGD in a lot of test cases, which made our work more solid.

I would also like to express my gratitude to Professor Konstantina Trivisa and Professor Maria Cameron for serving on my defense committee. Their help in my course study for PDEs and numerical analysis is also an essential building block of this thesis.

Table of Contents

Acknowledgements	ii
Table of Contents	iv
List of Tables	vi
List of Figures	vii
Chapter 1: Introduction	1
Chapter 2: Flocking Dynamics of Single- and Multi-Species	4
2.1 Introduction to (single-species) flocking dynamics	4
2.1.1 Particle description: the Cucker-Smale model	4
2.1.2 From particle to kinetic description	10
2.1.3 Hydrodynamic description of flocking	12
2.2 Hydrodynamic alignment of multi-species	17
2.2.1 Multi-species dynamics – description and main results	18
2.2.2 Graph connectivity – weighted Poincaré inequalities	22
2.2.3 Flocking of strong solutions	24
2.2.4 Boundedness of velocities	29
Chapter 3: Simulation Methods for Collective Dynamics	35
3.1 Particle simulations of flocking dynamics	36
3.2 Hydrodynamic simulations – numerical methods for conservation laws	41
3.2.1 Finite volume methods: upwind vs central	45
3.2.2 A brief review of second order central schemes	47
3.3 Revisiting the smooth van Albada slope limiter	52
3.3.1 The second order accuracy	53
3.3.2 The non-oscillatory property	54
3.4 Numerical examples	61
3.4.1 Conservation laws	61
3.4.2 Incompressible Euler equations	69
3.4.3 Hydrodynamic alignment of single-species	70
Chapter 4: Novel Swarm-Based Gradient Descent Method	75
4.1 The Swarm-Based Gradient Descent (SBGD) algorithm	76
4.1.1 Why communication is important	80

4.1.2	What is different about SBGD	84
4.2	Implementation of the SBGD_{pq} algorithm	86
4.2.1	Communications and mass transition	87
4.2.2	Time stepping – backtracking line search	87
4.2.3	SBGD_{pq} pseudocode	90
4.3	A general outlook	92
4.4	Convergence and error analysis	93
4.4.1	Convergence to a band of local minima	94
4.4.2	Flatness and convergence rate	98
4.5	Numerical results	101
4.5.1	One-dimensional problems	101
4.5.2	Two-dimensional problems	103
4.5.3	20-dimensional problems	112
Chapter 5:	Conclusions	122
5.1	Summary	122
5.2	Future works	124
Appendix A:	Proof of Theorem 2.1.5	126
Appendix B:	Derivation of Multi-Species Hydrodynamics	129
Appendix C:	Improved Dispersion Bound for One-Dimensional Multi-Species	133

List of Tables

4.1.1	Success rates of SBGD, GD(h), GD(BT) and Adam methods for global optimization of (4.1.4), based on $m = 1000$ runs with uniformly generated initial data in $[-3, 3]$. Backtracking parameters (see Algorithm 1), $\lambda = 0.2$ and $\gamma = 0.9$.	82
4.1.2	Success rates of SBGD, GD(h), GD(BT) and Adam methods for global optimization of (4.1.4) based on $m = 1000$ runs of uniformly generated initial data in $[-3, -1]$. Backtracking parameters (see Algorithm 1), $\lambda = 0.2$ and $\gamma = 0.9$	83
4.5.1	Results of SBGD for 1D Ackley and Rastrigin functions, $m = 200$.	102
4.5.2	Shifted 1D Ackley, results by SBGD, $m = 200$.	103
4.5.3	Shifted 1D Ackley, results by GD(BT) (no communication), $m = 200$.	104
4.5.4	2D Ackley, success rates of SBGD vs. GD(BT), $m = 500$.	109
4.5.5	2D Rastrigin, success rates of SBGD vs. GD(BT), $m = 500$.	109
4.5.6	2D drop-wave, success rates of SBGD vs. GD(BT), $m = 500$.	110
4.5.7	Average error for 2D Rosenbrock function SBGD vs. GD(BT), $m = 500$ simulations equi-distributed at $[-4, -2]^2$.	110
4.5.8	SBGD _{pq} for shifted 2D Rastrigin function; $N = 50$ agents	111
4.5.9	SBGD _{pq} for 2D Drop-wave function; $N = 10$ agents	111
4.5.10	2D Rosenbrock, SBGD _{pq} with $\lambda = 0.3$.	116
4.5.11	2D Rosenbrock, SBGD _{pq} with $\lambda = 0.5$.	116
4.5.12	2D Rosenbrock, SBGD _{pq} with $\lambda = 0.7$.	116
4.5.13	Success rates of SBGD compared with GD(h), GD(BT) and Adam methods for global optimization of (4.5.2) ($B = 0$), based on $m = 1000$ runs with uniformly generated initial data in $[-3, 3]^2$.	117
4.5.14	Success rates of SBGD compared with GD(h), GD(BT) and Adam methods for global optimization of 2D Rastrigin (4.5.2) ($B = 0$) based on $m = 1000$ runs of uniformly generated initial data in $[-3, -1]^2$.	117
4.5.15	80% success rate with 20D Ackley function	117
4.5.16	70% success rate with 20D Rastrigin function	117
4.5.17	$\mathbb{E}[\mathbf{X}_{SOL} - \mathbf{x}_*]$ for 20-dimensional shifted Ackley. $m = 1000$.	118
4.5.18	$\mathbb{E}[\mathbf{X}_{SOL} - \mathbf{x}_*]$ for 20-dimensional shifted Rastrigin. $m = 1000$.	118
4.5.19	20D Rastrigin. Average loss with $m = 1000$ simulations.	119
4.5.20	20D Ackley. Average loss with $m = 1000$ simulations.	120
4.5.21	Errors of the expectation and the corrected solution in $d = 20$, $B = C = 0$.	120
4.5.22	Solutions for the shifted Ackley function by SBGD, $d = 20$, $m = 1000$.	121
4.5.23	Solutions for the shifted Rastrigin function by SGBD, $d = 20$, $m = 1000$.	121

List of Figures

3.1.1 2D Cucker-Smale, trajectories of $N = 20$ particles up to $T = 5$. Red lines: trajectories of particles; blue lines: trajectories of mass centers.	41
3.1.2 2D Cucker-Smale, history of velocity fluctuations. x -axis: computation time; y -axis: velocity fluctuations.	42
3.4.1 Linear advection, NT scheme, $N = 200$, periodic bc, $t = 2\pi$	62
3.4.2 Linear advection, semi-discrete central-upwind scheme, $N = 200$, periodic bc, $t = 2\pi$	63
3.4.3 Sod's problem-density, NT scheme, $N = 400$, $t = 0.2$	64
3.4.4 Sod's problem-density, semi-discrete central-upwind scheme, $N = 400$, $t = 0.2$	64
3.4.5 Osher-Shu problem-density, NT scheme, $N = 600$, $t = 1.8$	65
3.4.6 Osher-Shu problem-density, semidiscrete central-upwind scheme, $N = 600$, $t = 1.8$	66
3.4.7 Double Mach reflection-density, 5th order WENO, $N_x = 480$, $N_y = 120$, $t = 0.2$	67
3.4.8 Double Mach reflection-density, NT scheme, minmod limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$	67
3.4.9 Double Mach reflection-density, NT scheme, van Albada limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$	68
3.4.10 Double Mach reflection-density, semidiscrete central-upwind scheme, minmod limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$	68
3.4.11 Double Mach reflection-density, semidiscrete central-upwind scheme, van Albada limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$	68
3.4.12 Double shear layer, contours of vorticity. $N_x = N_y = 128$, $CFL = 0.5$, $t = 8$	71
3.4.13 Double shear layer, solutions of v along $x = \pi$, $N_x = N_y = 128$, $CFL = 0.5$, $t = 8$	72
3.4.14 1D single-species with fat-tail kernel, history of energy fluctuation up to $T = 5$. $N = 100$, $CFL = 0.7$	73
3.4.15 2D single-species with fat-tail kernel, history of energy fluctuation up to $T = 4$. $N_x = N_y = 40$, $CFL = 0.7$	74
3.4.16 1D single-species with local communication, history energy fluctuation up to $T = 5$. $N = 100$, $CFL = 0.7$	74
4.1.1 Plot of the objective function (4.1.4).	81
4.1.2 Basins of attraction for GD(0.01), Adam(0.1), and Adam(1.1) method.	83
4.1.3 Histograms of problem (4.1.4) by $m = 200$ experiments. Initial data is generated uniformly in $[-3, -1]$. Global minimum at $x_* = 1.5355$	84
4.5.1 Benchmark functions	101
4.5.2 1D Ackley: evolution of agents in one simulation. $B = C = 0$	104

4.5.3 1D Rastrigin: evolution of agents in one simulation. $B = C = 0$	105
4.5.4 Histograms of the shifted 1D Ackley function by 200 experiments. $B = 25$, $C = 5$. Global minimum $\min f_{\text{Ackley}} = 5$ is attained at $\mathbf{x}_* = 25$	106
4.5.5 Landscapes of benchmark functions in two dimensions	108

Chapter 1: Introduction

Swarm dynamics has been an interesting and important research topic over the last decades. It investigates the relation between the interactions of microscopic individuals and the collective behavior of large crowd. Swarm dynamics has been a powerful tool to describe many phenomena from various research fields, including physics [46, 68, 83, 84, 104, 105, 135, 150], chemistry [15, 122], biology [1, 10, 26, 31, 32, 45, 77, 103, 119, 129, 138], ecology [4, 6, 9, 14, 28, 56, 61, 113, 152], human society [40, 59, 60], etc. A typical example from the nature is the formation of ant colony. While each ant alone only has limited local view, a large group of ants, however, can accomplish highly complicated and socialized configurations, such as food foraging, clustering and nest building. Interestingly, the emergence of these higher-level structures turns out to be the consequence of communications between ants based on the propagation of pheromone. Another example is the flocking of birds, which is typically associated with migration. The individual birds have tendency to gather as a flock to forage or travel collectively. By sharing information, the entire flock would have greater chance to find reliable food sources and be better protected from predators. The occurrence of collective intelligence in nature, resulting from cross-individual interactions, has also inspired the development of artificial, mission-oriented swarm dynamics systems. Examples include swarm robotics [5, 36, 47, 108, 110, 115, 130, 131, 146, 153], numerical optimization algorithms [64, 73, 76, 82, 116, 133, 157] and their wide applications in big data analysis, image processing, scheduling, vehicle routing, internet of things, resource allocation, wireless network, natural language processing [2, 21, 23, 24, 40, 78, 97], etc. The research of this thesis focuses on the analysis of flocking-based swarm dynamics and on the design of communication-based optimization algorithms.

The first aspect of this thesis is to study the flocking dynamics of self-driven particles. The mathematical study of flocking was first started from the work by Viscek [151] and further inspired by the hydrodynamic approach of Toner and Tu [144]. In our discussion, we focus on the long time behavior of self-organized systems driven by Cucker-Smale dynamics [29]. This well-known model has been used to describe the flocking mechanism with birds, with language evolution, etc. The system comprises of interacting agents of which the velocities are dynamically adjusted according to the average heading with neighbors. An important outcome of such communications is the flocking of the crowd, which emerges on both particle and hydrodynamic levels. The hydrodynamic limit of the Cucker-Smale system is described by an Eulerian system, associated with which a canonical issue is that the pressure tensor lacks closure for general social networks. Hence the pressure-less configuration has been assumed in many existing discussions, e.g. [44, 58, 72]. In this work we will analyze the hydrodynamics with respect to arbitrary pressure.

The second aspect of our study focuses on the swarm-based methods for numerical optimization. These methods are developed to provide near-optimal solutions to complicated non-convex optimization problems within a relatively short computation time. The main idea of swarm-based algorithms is to employ a group of simple individuals to search the solution domain. These individuals not only explore their surroundings but also exchange information with each other. In this way, the swarm of agents can better perceive the global landscape of the target function. So far there have been a lot of swarm-based algorithms proposed, such as particle swarm optimization and simulated annealing. Most of them, however, lack rigorous proofs regarding convergence and the optimality of solutions. In the recent year, the *Consensus-Based Optimization* (CBO) algorithm was proposed [116] which is supported up by a more complete analytical framework among the others. The basic set-up includes a group of particles of which the motions are governed by the drift towards an average state, i.e. the ‘consensus’, and a stochastic diffusion. The elegance of CBO lies in the application of Laplace principle [8], which dictates the concentration of particles within a small neighborhood of global minimizer. Nevertheless, the

performance of CBO, as well as the other gradient-free algorithms, can be highly sensitive to the choice of system parameters. In our work, we apply the idea of consensus formation to improve the performance of classical gradient-decent methods. A new swarm-based algorithm is proposed of which the convergence is assured regardless of artificial parameters. Moreover, thanks to the novel mass-based communications, agents are allowed to dynamically adjust individual search modes according to their relative performance, leading to notable improvement over the original single-agent dynamics.

The thesis is outlined as follows. In Chapter 2 we discuss the flocking dynamics for single- and multi-species systems. Discussions focus on the dynamics driven by long-range communication kernels. In particular, we prove the hydrodynamic alignment for connected multi-species networks without assuming a specific pressure.

Chapter 3 deals with the simulation methods for collective dynamics. One option is to discretize the agent-based model, in which case one keeps track of the motion of each particle. Alternatively, one can also simulate the behavior of hydrodynamic model, which is represented in terms of nonlinear transport equations. Different simulation approaches are applied to demonstrate the Cucker-Smale alignment through numerical computations.

In Chapter 4, we propose a novel *Swarm-Based Gradient Descent* (SBGD) method for non-convex optimization. The algorithm employs a population of agents to explore the domain, with each agent conducting local gradient descent search. A key ingredient is to assign each agent with an individual mass, which evaluates the quality of agent and determines the subsequent search strategy. Communications between agents are characterized by mass transitions and, as iterations proceed, the swarm will gradually concentrate the mass on the optimal candidate. The convergence of SBGD is rigorously studied. A series of benchmark tests are presented to demonstrate the advantage of introducing communications.

Chapter 2: Flocking Dynamics of Single- and Multi-Species

In this chapter, we study the large time behavior driven by the Cucker-Smale dynamics. The model is well known for its application in describing self-organized phenomena associated with flocking, such as the herding of birds and the schooling of fish.

In Section 2.1 we give an introduction to the flocking dynamics of single-species systems, i.e. agents interact under the same communication protocol. The description of flocking is demonstrated at particle and hydrodynamic levels. Our discussions focus on the dynamics driven by long-range communications with ‘fat-tails’, major flocking results are introduced.

Section 2.2 is the main contribution of this chapter. The discussion of flocking is extended to multi-species, where communication patterns between different clusters (species) may differ. Particularly, we are interested in the corresponding hydrodynamics describing the behavior of large crowd. Difficulty in analysis arises from the lack of closure for pressure. The innovation of our work is to prove the hydrodynamic alignment for multi-species with respect to *arbitrary* entropic pressure, based on graph connectivity and a weak dispersion bound. The results are joint work with Eitan Tadmor and have been published in [95].

2.1 Introduction to (single-species) flocking dynamics

2.1.1 Particle description: the Cucker-Smale model

Consider an N –particle interacting system consisting of identical particles, each of which is identified with a position/velocity pair, $(\mathbf{x}_i(t), \mathbf{v}_i(t)) \in \mathbb{R}^{2d}$, $1 \leq i \leq N$. The starting point of our discussion is the celebrated Cucker-Smale model [29, 30], which governs the particle motions

with the following equations,

$$\begin{cases} \frac{d}{dt}\mathbf{x}_i(t) = \mathbf{v}_i(t) \\ \frac{d}{dt}\mathbf{v}_i(t) = \frac{1}{N} \sum_{1 \leq j \leq N} \phi(\mathbf{x}_i(t), \mathbf{x}_j(t))(\mathbf{v}_j(t) - \mathbf{v}_i(t)) \end{cases} . \quad (2.1.1)$$

The interactions of particles are characterized by a non-negative symmetric function, ϕ , also known as the *communication kernel*,

$$\phi(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}', \mathbf{x}) \geq 0.$$

We say the system (2.1.1) is a '*single-species*' model to emphasize that the interactions between each pair of particles are characterized by a *uniform* communication kernel ϕ . In contrast, a '*multi-species*' model allows distinct communications for particles from different clusters (species), details will be discussed in Section 2.2.

We study the '*flocking*' behavior of single-species. In general, the flocking of particles indicates the occurrence of the two main features: (i) the spatial diameter, $\max_{i,j} |\mathbf{x}_j(t) - \mathbf{x}_i(t)|$, remains uniformly bounded in time and thus defining the 'flock', and (ii) the flock of particles is travelling with a bulk mean velocity asymptotically in time, $\mathbf{v}_i(t) \approx \bar{\mathbf{v}} \equiv \text{const}$. The second feature is known as the *alignment* dynamics.

To this end, we consider the bi-particle interactions which are decreasing functions of distance,

$$\phi(\mathbf{x}, \mathbf{x}') = \phi(|\mathbf{x} - \mathbf{x}'|), \quad \phi(\cdot) \text{ is decreasing.} \quad (2.1.2)$$

In particular, we focus on the dynamics driven by a long-range communication kernel that admits a *Pareto-type fat-tail*,

$$\phi(r) \gtrsim (1+r)^{-\theta}, \quad \theta > 0, \quad (2.1.3)$$

As shown in [51, 57, 143], the fat-tail decay plays a key role in the flocking of strong solutions.

A prototype example of fat-tail kernel was proposed by Cucker and Smale [29, 30]

$$\phi(r) = \frac{C}{(1 + r^2)^{\frac{\theta}{2}}}, \quad C > 0, \quad \theta > 0.$$

Such kernel is motivated by the reasoning that the strength of pairwise interactions is inversely dependent on the distance between two agents. Another important protocol of metric-based communication is the class of singular kernels (e.g. [17, 35, 99, 114, 118, 125, 126, 127]),

$$\phi(r) = \frac{1}{r^\beta}, \quad d < \beta < d + 2,$$

which emphasizes the near-by neighbors more strongly, $r \ll 1$, over those farther away.

We point out that a variety of communication protocols, which will not be analyzed here, have also been discussed in a large body of literature. Examples include, but not limited to, the short range kernels [141] where $\phi(\mathbf{x}, \mathbf{x}')$ is compactly supported within the distance $|\mathbf{x} - \mathbf{x}'| \leq R_0$; the heterophilous dynamics [101] where $\phi(r)$ is increasing over its support, leading to the tendency to attract diverse groups; the topological-based kernels (e.g. [7, 128]) where $\phi(\mathbf{x}, \mathbf{x}') = \varphi(\mu(\mathcal{C}(\mathbf{x}, \mathbf{x}')))$ is determined by the density of the crowd in an intermediate domain $\mathcal{C}(\mathbf{x}, \mathbf{x}')$ enclosed between $\mathbf{x}$ and $\mathbf{x}'$,

$$\mu(\mathbf{x}, \mathbf{x}') = \frac{1}{N} \# \{j : \mathbf{x}_j \in \mathcal{C}(\mathbf{x}, \mathbf{x}')\}.$$

In particular, a *non-symmetric* topological kernel was studied in [100] where the domain of communication $\mathcal{C}$ is taken to an R -ball centered at $\mathbf{x}$, $\phi(\mathbf{x}, \mathbf{x}') = \varphi(|\mathbf{x} - \mathbf{x}'|)/\mu(B_R(\mathbf{x}))$. We refer the details to respective references.

Now we investigate the large time behavior of solutions to system (2.1.1), (2.1.2). To this end, we introduce the center of mass $(\mathbf{x}_c(t), \mathbf{v}_c(t))$,

$$\mathbf{x}_c(t) = \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j(t), \quad \mathbf{v}_c(t) = \frac{1}{N} \sum_{j=1}^N \mathbf{v}_j(t).$$

It is clear that $\frac{d}{dt}\mathbf{x}_c(t) = \mathbf{v}_c(t)$. A direct implication of (2.1.1) is the conservation of total momentum, $\sum_{j=1}^N \mathbf{v}_j(t)$. Indeed, the symmetry $\phi(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_j, \mathbf{x}_i)$ implies

$$\frac{d}{dt}\left(\sum_{i=1}^N \mathbf{v}_i\right) = \sum_{1 \leq i, j \leq N} \phi(\mathbf{x}_i, \mathbf{x}_j)(\mathbf{v}_j - \mathbf{v}_i) = 0.$$

The conservation of momentum indicates that the mean velocity, $\mathbf{v}_c$, remains constant in time, and the center, $\mathbf{x}_c$, moves along a straight line,

$$\mathbf{v}_c(t) = \mathbf{v}_c(0), \quad \mathbf{x}_c(t) = \mathbf{x}_c(0) + t\mathbf{v}_c(0).$$

To analyze the difference of individual behaviors from the center of system, the two auxiliary functions are introduced which measure the fluctuations in positions and velocities [51],

$$X(t) = \sum_{i=1}^N |\mathbf{x}_i(t) - \mathbf{x}_c(t)|^2, \quad V(t) = \sum_{i=1}^N |\mathbf{v}_i(t) - \mathbf{v}_c|^2, \quad (2.1.4)$$

subject to initial data $(X(0), V(0)) = (X_0, V_0)$. We denote the spatial diameter by

$$D(t) := \max_{1 \leq i, j \leq N} |\mathbf{x}_i(t) - \mathbf{x}_j(t)|.$$

It is clear that $D(t) \leq \sqrt{2X(t)}$. The strength of communication is bounded from below with

$$\min_{1 \leq i, j \leq N} \phi(|\mathbf{x}_i(t) - \mathbf{x}_j(t)|) \geq \phi(D(t)) \geq \phi(\sqrt{2X(t)}). \quad (2.1.5)$$

We start with the following estimate on the velocity fluctuations.

Lemma 2.1.1 *Let $(\mathbf{x}_i(t), \mathbf{v}_i(t))$ be a solution of the system (2.1.1), (2.1.2). Then we have*

$$V(t) \leq V_0 \exp\left\{-2 \int_0^t \phi(D(s)) ds\right\}. \quad (2.1.6)$$

Proof. The proof follows the lines of [51, Lemma 2.2]. By the definition of $\mathbf{v}_c$, we rewrite $V(t)$ as

$$V(t) = \sum_{i=1}^N |\mathbf{v}_i(t) - \mathbf{v}_c|^2 = \sum_{j=1}^N |\mathbf{v}_j|^2 - N|\mathbf{v}_c|^2.$$

Since $\mathbf{v}_c$ is independent of time, by the second equation of (2.1.1) and the symmetry of ϕ , we have

$$\begin{aligned} \frac{d}{dt}V(t) &= \sum_{i=1}^N 2\mathbf{v}_i \cdot \frac{d\mathbf{v}_i}{dt} \\ &= \frac{2}{N} \sum_{1 \leq i, j \leq N} \phi(|\mathbf{x}_i - \mathbf{x}_j|) \mathbf{v}_i \cdot (\mathbf{v}_j - \mathbf{v}_i) \\ &= \frac{2}{N} \sum_{1 \leq i, j \leq N} \phi(|\mathbf{x}_i - \mathbf{x}_j|) \mathbf{v}_j \cdot (\mathbf{v}_i - \mathbf{v}_j) \\ &= -\frac{1}{N} \sum_{1 \leq i, j \leq N} \phi(|\mathbf{x}_i - \mathbf{x}_j|) |\mathbf{v}_i - \mathbf{v}_j|^2 \end{aligned}$$

Furthermore, by the lower bound (2.1.5),

$$\begin{aligned} \frac{d}{dt}V(t) &\leq -\frac{\phi(D(t))}{N} \sum_{1 \leq i, j \leq N} |\mathbf{v}_i - \mathbf{v}_j|^2 \\ &= -2\phi(D(t)) \sum_{i=1}^N |\mathbf{v}_i - \mathbf{v}_c|^2 \\ &= -2\phi(D(t))V(t) \end{aligned}$$

The equation at the second line is obtained by writing $\mathbf{v}_i - \mathbf{v}_j = \mathbf{v}_i - \mathbf{v}_c - (\mathbf{v}_j - \mathbf{v}_c)$ and noticing that $\sum_{i=1}^N (\mathbf{v}_i - \mathbf{v}_c) = 0$. Integration of the above inequality yields the estimate (2.1.6). $\square$

Assuming the communication kernel has a fat-tail (2.1.3), the estimate (2.1.6) reduces to

$$V(t) \lesssim V_0 \exp\left\{-\int_0^t (1 + D(s))^{-\theta} ds\right\}. \quad (2.1.7)$$

It is noticed that the decay of $V(t)$ is dependent on the growth of diameter $D(t)$. The following lemma shows a quadratic growth rate of the fluctuations in space.

Lemma 2.1.2 *Let $(\mathbf{x}_i(t), \mathbf{v}_i(t))$ be a solution of the system (2.1.1), (2.1.2). Then we have*

$$X(t) \leq 2X_0 + 2V_0 t^2. \quad (2.1.8)$$

Proof. The proof follows the lines of [51, Lemma 2.3]. We apply the Cauchy-Schwartz's inequality,

$$\begin{aligned} \frac{d}{dt} X(t) &= \sum_{j=1}^N 2(\mathbf{x}_i - \mathbf{x}_v) \left(\frac{d\mathbf{x}_i}{dt} - \frac{d\mathbf{x}_c}{dt} \right) \\ &\leq 2 \sqrt{\sum_{i=1}^N |\mathbf{x}_i - \mathbf{x}_c|^2} \sqrt{\sum_{i=1}^N |\mathbf{v}_i - \mathbf{v}_c|^2} \\ &= 2\sqrt{X(t)}\sqrt{V(t)} \\ &\leq 2\sqrt{X(t)}\sqrt{V_0} \exp\left\{-\int_0^t \phi(D(s))ds\right\} \end{aligned}$$

Integration in time yields

$$\sqrt{X(t)} \leq \sqrt{X_0} + \sqrt{V_0} \int_0^t e^{-\int_0^s \phi(D(\tau))d\tau} ds \leq \sqrt{X_0} + \sqrt{V_0} t.$$

By squaring both sides of the inequality one obtains (2.1.8). $\square$

The lemma implies a linear growth in the spatial diameter, $D(t) \lesssim 1+t$. Substitute this into (2.1.7), one recovers a fractional exponential decay in $V(t)$ for $\theta < 1$ and a polynomial decay for $\theta = 1$,

$$V(t) \lesssim V_0 \times \begin{cases} e^{-\nu_1 t^{1-\theta}}, & \theta \in [0, 1), \quad \nu_1 = \nu_1(X_0, V_0, \theta) \\ (1+t)^{-\nu_2}, & \theta = 1, \quad \nu_2 = \nu_2(X_0, V_0) \end{cases}.$$

In fact, for the case $\theta < 1$, one can obtain a full exponential decay rate through a bootstrapping argument [51]. The improved result is stated as follows.

Theorem 2.1.3 ([51, Theorem 2.4]) *Let $(\mathbf{x}_i(t), \mathbf{v}_i(t))$ be a solution to (2.1.1), (2.1.3). Then the following statements are true,*

(i) There exists a positive constant, C , depending on X_0, V_0, θ , such that

$$|\mathbf{x}_i(t) - \mathbf{x}_c| \lesssim |\mathbf{x}_i(0) - \mathbf{x}_c| + C. \quad (2.1.9)$$

(ii) There exists constants, $\nu_1, \nu_2 > 0$, such that

$$V(t) \lesssim V_0 \times \begin{cases} e^{-\nu_1 t}, & \theta \in (0, 1), & \nu_1 = \nu_1(X_0, V_0, \theta) \\ (1+t)^{-\nu_2}, & \theta = 1, & \nu_2 = \nu_2(X_0, V_0) \end{cases}. \quad (2.1.10)$$

Theorem 2.1.3 verifies the two main features of flocking for $\theta \leq 1$: the diameter $\max_{i,j} |\mathbf{x}_i - \mathbf{x}_j|$ remains uniformly bounded, and the flock of particles is travelling with a bulk mean velocity asymptotically, $\mathbf{v}_i(t) \rightarrow \mathbf{v}_c(0)$ as $t \rightarrow \infty$. For $\theta > 1$, the flocking may also occur under certain initial configurations. We refer the discussions to [30].

2.1.2 From particle to kinetic description

When there is a sufficiently large number of particles, it is not economical to keep track of the motion of each particle. Instead, it would be meaningful to study the mean-field limit of the particle system (2.1.1), namely, a kinetic (mesoscopic) description in analogy with the Vlasov equation in plasma and astrophysics.

A formal derivation of kinetic model was presented in [51] based on the BBGKY hierarchy [11]. Here we proceed with an alternative approach based on the use of empirical distribution. The derivation follows the lines of [58]. We define the empirical distribution, $f(\mathbf{x}, \mathbf{v}, t)$, associated to the system (2.1.1),

$$f(t, \mathbf{x}, \mathbf{v}) = \frac{1}{N} \sum_{i=1}^N \delta_{\mathbf{x}_i}(\mathbf{x}) \otimes \delta_{\mathbf{v}_i}(\mathbf{v}), \quad (t, \mathbf{x}, \mathbf{v}) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}_+. \quad (2.1.11)$$

The distribution density function f describes the probability of finding a particle with position $\mathbf{x}$ and velocity $\mathbf{v}$ at time t . For simplicity of notation, we will omit the time variable t in the

distribution functions whenever the time-dependence is clear by the context. We test $\partial_t f$ against an arbitrary smooth function,

$$\begin{aligned}
\iint \partial_t f(t, \mathbf{x}, \mathbf{v}) \varphi(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} &= \frac{d}{dt} \left[\frac{1}{N} \sum_{i=1}^N \varphi(\mathbf{x}_i(t), \mathbf{v}_i(t)) \right] \\
&= \frac{1}{N} \sum_{i=1}^N [\dot{\mathbf{x}}_i \cdot \nabla_{\mathbf{x}} \varphi(\mathbf{x}_i, \mathbf{v}_i) + \dot{\mathbf{v}}_i \cdot \nabla_{\mathbf{v}} \varphi(\mathbf{x}_i, \mathbf{v}_i)] \quad (2.1.12) \\
&= \frac{1}{N} \sum_{i=1}^N [\mathbf{v}_i \cdot \nabla_{\mathbf{x}} \varphi(\mathbf{x}_i, \mathbf{v}_i) + F^i \cdot \nabla_{\mathbf{v}} \varphi(\mathbf{x}_i, \mathbf{v}_i)],
\end{aligned}$$

the alignment of particles is driving by the forcing term,

$$F^i = \frac{d}{dt} \mathbf{v}_i(t) = \frac{1}{N} \sum_{j=1}^N \phi(|\mathbf{x}_i - \mathbf{x}_j|) (\mathbf{v}_j - \mathbf{v}_i) = L[f],$$

where $L[f](\mathbf{x}_i, \mathbf{v}_i) := \iint \phi(|\mathbf{x}_i - \mathbf{x}_j|) (\mathbf{v}' - \mathbf{v}_i) f(\mathbf{x}', \mathbf{v}') d\mathbf{x}' d\mathbf{v}'$. Integrate (2.1.12) by part, we obtain

$$\begin{aligned}
\iint \partial_t f(t, \mathbf{x}, \mathbf{v}) \varphi(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} &= \iint [\mathbf{v} \cdot \nabla_{\mathbf{x}} \varphi(\mathbf{x}, \mathbf{v}) + L[f] \cdot \nabla_{\mathbf{v}} \varphi(\mathbf{x}, \mathbf{v})] f(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} \\
&= - \iint [\mathbf{v} \cdot \nabla_{\mathbf{x}} f + \nabla_{\mathbf{v}} \cdot (f L[f])] \varphi d\mathbf{x} d\mathbf{v}.
\end{aligned}$$

Since the test function φ is arbitrary, one obtains a Vlasov-type mean-field equation governing the evolution of one-particle distribution $f(\mathbf{x}, \mathbf{v}, t)$,

$$\begin{aligned}
\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f + \nabla_{\mathbf{v}} \cdot (f L[f]) &= 0, \\
L[f](t, \mathbf{x}, \mathbf{v}) &= \int_{\mathbb{R}^{2d}} \phi(\mathbf{x}, \mathbf{x}') (\mathbf{v}' - \mathbf{v}) f(t, \mathbf{x}', \mathbf{v}') d\mathbf{x}' d\mathbf{v}'. \quad (2.1.13)
\end{aligned}$$

The bi-linear term $f L[f]$ characterizes the interactions between particles. We finish the formal derivation of the kinetic description.

2.1.3 Hydrodynamic description of flocking

To investigate the collective behaviors of a large crowd, $N \rightarrow \infty$, we will study the hydrodynamic limit of Cucker-Smale dynamics, which can be formally derived by taking the $\mathbf{v}$ -moments of the kinetic description (2.1.13). We first define the hydrodynamic variables: the mass and the momentum,

$$\rho(t, \mathbf{x}) := \lim_{N \rightarrow \infty} \int_{\mathbb{R}^d} f(t, \mathbf{x}, \mathbf{v}) d\mathbf{v} \quad \text{and} \quad \rho \mathbf{u}(t, \mathbf{x}) = \lim_{N \rightarrow \infty} \int_{\mathbb{R}^d} \mathbf{v} f(t, \mathbf{x}, \mathbf{v}) d\mathbf{v}.$$

The large crowd limits are assumed to exist and were justified, at least, for mildly singular kernels, [16, 19, 44, 51, 106]. Multiply the equation (2.1.13) by 1 and $\mathbf{v}$ and integrate over the velocity space $\mathbf{v} \in \mathbb{R}^d$, one recovers the hydrodynamic model for a single-species system

$$\begin{cases} \partial_t \rho + \nabla_{\mathbf{x}} \cdot (\rho \mathbf{u}) = 0, \\ \partial_t (\rho \mathbf{u}) + \nabla_{\mathbf{x}} \cdot (\rho \mathbf{u} \otimes \mathbf{u} + \mathbb{P}) = \mathbf{A}(\rho, \mathbf{u}), \end{cases} \quad (t, \mathbf{x}) \in (\mathbb{R}_t, \Omega), \quad (2.1.14)$$

where nonlocal source term,

$$\mathbf{A}(\rho, \mathbf{u}) = \int_{\mathbb{R}^d} \phi(|\mathbf{x} - \mathbf{x}'|) (\mathbf{u}(t, \mathbf{x}') - \mathbf{u}(t, \mathbf{x})) \rho(t, \mathbf{x}) \rho(t, \mathbf{x}') d\mathbf{x}' d\mathbf{v}' \quad (2.1.15)$$

characterizes the alignment interactions. The pressure tensor, $\mathbb{P}$, is given by the second $\mathbf{v}$ -moment of the kinetic distribution $f(\mathbf{x}, \mathbf{v}, t)$,

$$\mathbb{P}(t, \mathbf{x}) = \int_{\mathbb{R}^d} (\mathbf{v} - \mathbf{u}) \otimes (\mathbf{v} - \mathbf{u}) f(t, \mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v}. \quad (2.1.16)$$

Integrate the equations (2.1.13) over the spatial variable $\mathbf{x} \in \mathbb{R}^d$, one obtains the conservation of the total mass, $M := \int \rho d\mathbf{x}$, and the total momentum, $\mathbf{m}_1 := \int \rho \mathbf{u} d\mathbf{x}$.

We notice that the pressure tensor $\mathbb{P}$ lacks closure. Indeed, as the hydrodynamic model only includes a finite number of equations, the information of the highest $\mathbf{v}$ -moment is inevitably

truncated. Thus, strictly speaking, the model (2.1.14) is a mesoscopic description unless the closure is imposed from the ‘outside’. In particle physics, the closure problem is resolved by imposing the canonical Maxwellian distribution which prefers a state of thermal equilibrium

$$f_{\{\rho, \mathbf{u}, T\}}(t, \mathbf{x}, \mathbf{v}) = \frac{\rho}{(2\pi T)^d} e^{-\frac{|\mathbf{v} - \mathbf{u}|^2}{2T}},$$

with ρ , $\mathbf{u}$, T being the density, velocity and pressure respectively. The special case of vanishing temperature, $T \downarrow 0$, leads to the *mono-kinetic* closure, $\mathbb{P} \equiv 0$, associated with which $f(\mathbf{x}, \mathbf{v}, t) = \rho \delta(|\mathbf{v} - \mathbf{u}|)$ [44, 72]. Other examples include the entropic-based closure with measure-value [13], and the isothermal closure (constant temperature T), $\mathbb{P} = \rho \mathbb{I}_{d \times d}$, [75]. In the current setup of general ‘social particles’, however, the question of closure is left open.

Despite the lack of closure of higher moments, an estimate on the flocking mechanism can still be proved at the hydrodynamic scale. To this end, we define the total energy given by the second moment

$$\rho E(t, \mathbf{x}) := \int_{\mathbb{R}^d} \frac{|\mathbf{v}|^2}{2} f(t, \mathbf{x}, \mathbf{v}) d\mathbf{v}.$$

By writing $|\mathbf{v}|^2 = |\mathbf{v} - \mathbf{u}|^2 + |\mathbf{u}|^2 + 2\mathbf{u} \cdot \mathbf{v}$ and noticing that $\int (\mathbf{v} - \mathbf{u}) f d\mathbf{v} = 0$, the total energy is decomposed into the sum of the kinetic energy and the internal energy,

$$\rho E = \frac{\rho |\mathbf{u}|^2}{2} + \rho e, \quad \rho e(t, \mathbf{x}) := \int \frac{|\mathbf{v} - \mathbf{u}(t, \mathbf{x})|^2}{2} f(\mathbf{x}, \mathbf{v}, t) d\mathbf{v}.$$

Integrate the equation (2.1.13) against $\frac{|\mathbf{v}|^2}{2}$, one obtains the energy equation

$$\begin{aligned} \partial_t(\rho E) + \nabla_{\mathbf{x}} \cdot (\rho E \mathbf{u} + \mathbb{P} \mathbf{u} + \mathbf{q}) = \\ - \int \phi(\mathbf{x}, \mathbf{x}') (2E(t, \mathbf{x}) - \mathbf{u}(t, \mathbf{x}) \cdot \mathbf{u}(t, \mathbf{x}')) \rho(t, \mathbf{x}) \rho(t, \mathbf{x}') d\mathbf{x}' \end{aligned} \quad (2.1.17)$$

where $\mathbf{q} := \int \frac{|\mathbf{v} - \mathbf{u}|^2}{2} (\mathbf{v} - \mathbf{u}) f d\mathbf{v}$ is the heat flux. The source term on the right is known as the *enstrophy*. Integration of equation (2.1.17) yields the decay in the global energy, $\mathcal{E}(t) :=$

$$\int \rho E \, d\mathbf{x},$$

$$\begin{aligned} \frac{d}{dt} \mathcal{E}(t) &= - \iint \phi(|\mathbf{x} - \mathbf{x}'|) (2E(\mathbf{x}) - \mathbf{u}(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x}')) \rho(\mathbf{x}) \rho(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= - \iint \phi(|\mathbf{x} - \mathbf{x}'|) (2E(\mathbf{x}') - \mathbf{u}(\mathbf{x}') \cdot \mathbf{u}(\mathbf{x})) \rho(\mathbf{x}) \rho(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= - \iint \phi(|\mathbf{x} - \mathbf{x}'|) (E(\mathbf{x}) + E(\mathbf{x}') - \mathbf{u}(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x}')) \rho(\mathbf{x}) \rho(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= - \iint \phi(|\mathbf{x} - \mathbf{x}'|) \left(\frac{|\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x}')|^2}{2} + e(\mathbf{x}) + e(\mathbf{x}') \right) \rho(\mathbf{x}) \rho(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \end{aligned} \quad (2.1.18)$$

The analysis of flocking behavior is based on estimation of the *energy fluctuations*,

$$\delta \mathcal{E}(t) = \iint \left(\frac{|\mathbf{u}(t, \mathbf{x}) - \mathbf{u}(t, \mathbf{x}')|^2}{2} + e(t, \mathbf{x}) + e(t, \mathbf{x}') \right) \rho(t, \mathbf{x}) \rho(t, \mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}'.$$

This functional incorporates the fluctuations velocity at the macroscopic scale, as given by the first term, and at the microscopic scale, as measured by the internal energies.

We have the following theorem.

Theorem 2.1.4 Assume $(\rho, \mathbf{u}, \mathbb{P}) \in (L^\infty \cap L^1_+(\mathbb{R}^d)) \times W^{1,\infty}(\mathbb{R}^d) \times W^{1,\infty}(\mathbb{R}^d)$ is a strong solution of the system (2.1.14)–(2.1.15) subject to compactly supported initial data $(\rho_0, \mathbf{u}_0, \mathbb{P}_0)$. Let $D(t)$ denote the diameter of the support,

$$D(t) := \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{S}(t)} |\mathbf{x} - \mathbf{y}|, \quad \mathcal{S}(t) := \text{supp}\{\rho(t, \cdot)\}.$$

Then the energy fluctuations $\delta \mathcal{E}(t)$ satisfies the decay bound

$$\delta \mathcal{E}(t) \leq \delta \mathcal{E}_0 \cdot \exp\left\{-2M \int_0^t \phi(D(s)) \, ds\right\}. \quad (2.1.19)$$

Here M is the total mass, $M = \int \rho \, d\mathbf{x} \equiv \text{const.}$

Proof. The proof follows the lines of [51, 141, 142]. We represent the functional $\delta \mathcal{E}(t)$ in terms

of the mass, M , and the momentum, $\mathbf{m}_1$,

$$\delta\mathcal{E}(t) = 2M\mathcal{E}(t) - |\mathbf{m}_1|^2.$$

We use (2.1.18) to find

$$\begin{aligned} \frac{d}{dt}\delta\mathcal{E}(t) &= -2M \iint \phi(|\mathbf{x} - \mathbf{x}'|) \left(\frac{|\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x}')|^2}{2} + e(\mathbf{x}) + e(\mathbf{x}') \right) \rho(\mathbf{x}) \rho(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &\leq -2M\phi(D(t)) \iint \left(\frac{|\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x}')|^2}{2} + e(\mathbf{x}) + e(\mathbf{x}') \right) \rho(\mathbf{x}) \rho(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= -2M\phi(D(t))\delta\mathcal{E}(t). \end{aligned}$$

Grownwall's inequality yields the desired results. $\square$

The non-negative of internal energy implies

$$\begin{aligned} \delta\mathcal{E}(t) &\geq \iint \frac{|\mathbf{u}(t, \mathbf{x}) - \mathbf{u}(t, \mathbf{x}')|^2}{2} \rho(t, \mathbf{x}) \rho(t, \mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= M \iint |\mathbf{u}(t, \mathbf{x}) - \mathbf{u}_\infty|^2 \rho(t, \mathbf{x}) \, d\mathbf{x}, \end{aligned} \tag{2.1.20}$$

where the *average velocity*, $\mathbf{u}_\infty := \frac{\int \rho \mathbf{u} \, d\mathbf{x}}{\int \rho \, d\mathbf{x}}$, is time-independent. Combining (2.1.19) and (2.1.20), we observe that the alignment of velocity,

$$\int |\mathbf{u}(t, \mathbf{x}) - \mathbf{u}_\infty|^2 \rho(t, \mathbf{x}) \, d\mathbf{x} \rightarrow 0,$$

emerges if the diameter $D(t)$ satisfies the condition

$$\int_0^t \phi(D(t)) \, dt \rightarrow \infty, \quad t \rightarrow \infty.$$

The growth rate of $D(t)$ is bounded by the pointwise velocity fluctuations,

$$\frac{d}{dt}D(t) \leq \delta V(t), \quad \delta V(t) := \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{S}_0} |\mathbf{u}(t, \mathbf{x}) - \mathbf{u}(t, \mathbf{y})|.$$

Estimation of $\delta V(t)$ with respect to general pressure remains an open problem. However, in the particular case of mono-kinetic closure, $\mathbb{P} \equiv 0$, we have the following estimates.

Theorem 2.1.5 *Let $(\rho(t, \cdot), \mathbf{u}(t, \cdot)) \in (L^\infty \cap L^1_+(\mathbb{R}^d)) \times W^{1,\infty}(\mathbb{R}^d)$ be a strong solution to the system (2.1.14)–(2.1.15) coupled with vanishing pressure, $\mathbb{P} \equiv 0$, subject to compactly supported initial data, $(\rho_0, \mathbf{u}_0)$, with finite velocity fluctuations,*

$$\delta V_0 := \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{S}_0} |\mathbf{u}_0(\mathbf{x}) - \mathbf{u}_0(\mathbf{y})| < \infty, \quad \mathcal{S}_0 := \text{supp}\{\rho_0(\cdot)\}.$$

Assume the associated communication kernel, $\phi(r)$, satisfies a Pareto-type fat-tail condition

$$\phi(r) \gtrsim (1+r)^{-\theta}, \quad \theta \in (0, 1].$$

Then the following holds,

(i) The spatial diameter is bounded by

$$D(t) \leq \begin{cases} D_0 + t\delta V_0, & \theta = 1 \\ D_\infty, & \theta \in (0, 1) \end{cases}. \quad (2.1.21)$$

The constant, $D_\infty > 0$, depends on $M, D_0, \delta V_0, \theta$.

(ii) There exists constants, $\nu_1, \nu_2 > 0$, such that

$$\delta V(t) \lesssim \delta V_0 \times \begin{cases} (1+t)^{-\nu_1}, & \theta = 1, \quad \nu_1 = M(\max\{1 + D_0, \delta V_0\})^{-1} \\ e^{-\nu_2 t}, & \theta \in (0, 1), \quad \nu_2 = M\phi(D_\infty). \end{cases}. \quad (2.1.22)$$

The proof of the theorem follows the lines of [57, 58]. Details are given in Appendix A.

We obtain an immediate corollary.

Corollary 2.1.6 *Consider the single-species hydrodynamic system (2.1.14) associated with a fat-tail kernel (2.1.3). Assume $(\rho(t, \cdot), \mathbf{u}(t, \cdot))$ is a strong solution subject to compactly supported*

initial data with finite velocity fluctuations, $\delta V_0 < \infty$. Then the solution exhibits a large time flocking behavior where velocity is asymptotically aligned along the average, $\mathbf{u}_\infty := \frac{\int \rho \mathbf{u} \, d\mathbf{x}}{\int \rho \, d\mathbf{x}}$,

$$\int |\mathbf{u}(t, \mathbf{x}) - \mathbf{u}_\infty|^2 \rho(t, \mathbf{x}) \, d\mathbf{x} \lesssim \int |\mathbf{u}_0(\mathbf{x}) - \mathbf{u}_\infty|^2 \rho_0(\mathbf{x}) \, d\mathbf{x} \cdot \begin{cases} (1+t)^{-\nu_3}, & \theta = 1 \\ e^{-\nu_4 t}, & \theta \in (0, 1) \end{cases}, \quad (2.1.23)$$

where $\nu_3 = 2M(\max\{1 + D_0, \delta V_0\})^{-1}$, $\nu_4 = 2M\phi(D_\infty)$.

Proof. Substitute (2.1.21) into (2.1.19), we obtain

$$\delta \mathcal{E}(t) \lesssim \delta \mathcal{E}(0) \cdot \begin{cases} (1+t)^{-\nu_3}, & \theta = 1 \\ e^{-\nu_4 t}, & \theta \in (0, 1) \end{cases},$$

The desired result is recovered by noticing that

$$\begin{aligned} \delta \mathcal{E}(t) &= \iint \frac{|\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x}')|^2}{2} \rho(\mathbf{x}) \rho(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= M \int |\mathbf{u}(t, \mathbf{x}) - \mathbf{u}_\infty|^2 \rho(t, \mathbf{x}) \, d\mathbf{x}, \end{aligned}$$

due to the vanishing internal energy, $\rho e = \frac{1}{2} \text{trace}(\mathbb{P}) \equiv 0$. □

2.2 Hydrodynamic alignment of multi-species

We extend the study of flocking dynamics to multi-species, where non-uniform communications are allowed. Discussions in this section are based on the joint paper with Eitan Tadmor [95], which is developed from the forerunner [142] regarding single-species hydrodynamics with pressure.

2.2.1 Multi-species dynamics – description and main results

2.2.1.1 Hydrodynamic description of multi-species

We study the large-time behavior of the multi-species (hydro-)dynamics

$$\begin{cases} \partial_t \rho_\alpha + \nabla_{\mathbf{x}} \cdot (\rho_\alpha \mathbf{u}_\alpha) = 0, \\ \partial_t (\rho_\alpha \mathbf{u}_\alpha) + \nabla_{\mathbf{x}} \cdot (\rho_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha + \mathbb{P}_\alpha) = \mathbf{A}_\alpha(\rho, \mathbf{u}), \end{cases} \quad (t, \mathbf{x}) \in (\mathbb{R}_+, \mathbb{R}^d), \quad (2.2.1)$$

subject to initial data $(\rho_\alpha, \mathbf{u}_\alpha, \mathbb{P}_\alpha)|_{t=0} = (\rho_{\alpha 0}, \mathbf{u}_{\alpha 0}, \mathbb{P}_{\alpha 0})$. The different species, tagged by the (possibly infinite) index-set $\mathcal{I} \ni \alpha$, are quantified by their density, $\rho_\alpha : \mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}_+$, momentum, $(\rho \mathbf{u})_\alpha : \mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}^d$, and pressure tensor, $\mathbb{P}_\alpha : \mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}^d \times \mathbb{R}^d$. The dynamics is driven by inter-species interactions due to *alignment*

$$\mathbf{A}_\alpha(\rho, \mathbf{u}) := \sum_{\beta \in \mathcal{I}} \int \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (\mathbf{u}_\beta(t, \mathbf{x}') - \mathbf{u}_\alpha(t, \mathbf{x})) \rho_\alpha(t, \mathbf{x}) \rho_\beta(t, \mathbf{x}') d\mathbf{x}', \quad (2.2.2)$$

dictated by an array of symmetric communications kernels, $\Phi = \{\phi_{\alpha\beta}\}$ where $\phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') = \phi_{\beta\alpha}(\mathbf{x}', \mathbf{x}) \geq 0$. Integrate equations (2.2.1) over $\mathbf{x} \in \mathbb{R}^d$ and sum over $\alpha \in \mathcal{I}$, one recovers the conservation of total mass, $M := \sum_{\alpha \in \mathcal{I}} \int \rho_\alpha d\mathbf{x}$, and total momentum, $\mathbf{m}_1 := \sum_{\alpha \in \mathcal{I}} \int \rho_\alpha \mathbf{u}_\alpha d\mathbf{x}$, of the overall crowd.

The prominent role of metric kernels enters when we assume that there exists a symmetric array of radial kernels $K := \{k_{\alpha\beta}(r)\}$ such that

$$\phi_{\alpha\beta}(\mathbf{x}', \mathbf{x}) \geq k_{\alpha\beta}(|\mathbf{x} - \mathbf{x}'|) = k_{\beta\alpha}(|\mathbf{x} - \mathbf{x}'|) \geq 0, \quad \alpha, \beta \in \mathcal{I}. \quad (2.2.3)$$

We use the notation $\Phi(\mathbf{x}, \mathbf{x}') \succeq K(|\mathbf{x} - \mathbf{x}'|)$ to abbreviate (2.2.3). The different species are distinguished by the protocol of their interactions with the rest of the crowd, as characterized by different communication kernels, $\phi_{\alpha\beta}$, between the species α and β . There are two extreme cases: when $\phi_{\alpha\beta} \equiv \phi$, one recovers the single-species hydrodynamics (2.1.14)–(2.1.15). When

$\phi_{\alpha\beta} = \delta_{\alpha\beta}\phi$, we end up with independent and identical copies of single-species. In the following discussions, we focus on the intermediate cases of *genuine* multi-species dynamics.

The multi-species system (2.2.1)–(2.2.3) corresponds to the large-crowd dynamics of the agent-based model [58], where different species, each of which consists of N_α agents with position/velocity $(\mathbf{x}_\alpha^i(t), \mathbf{v}_\alpha^i(t)) : \mathbb{R}_+ \mapsto \mathbb{R}^d \times \mathbb{R}^d$, are driven by the Cucker-Smale alignment

$$\begin{cases} \frac{d}{dt} \mathbf{x}_\alpha^i(t) = \mathbf{v}_\alpha^i(t), \\ \frac{d}{dt} \mathbf{v}_\alpha^i(t) = \sum_{\beta \in \mathcal{I}} \frac{1}{N_\beta} \sum_{j=1}^{N_\beta} \phi_{\alpha\beta}(\mathbf{x}_\beta^j(t), \mathbf{x}_\alpha^i(t)) (\mathbf{v}_\beta^j(t) - \mathbf{v}_\alpha^i(t)), \end{cases} \quad i = 1, 2, \dots, N_\alpha, \quad (2.2.4)$$

subject to initial data $(\mathbf{x}_\alpha^i(0), \mathbf{v}_\alpha^i(0)) = (\mathbf{x}_\alpha^{i0}, \mathbf{v}_\alpha^{i0})$. The passage from the agent-based to the hydrodynamic description is realized by moments of the empirical distribution (see Appendix B)

$$f_\alpha(t, \mathbf{x}, \mathbf{v}) := \frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} \delta_{\mathbf{x}_\alpha^i(t)} \otimes \delta_{\mathbf{v}_\alpha^i(t)}, \quad (t, \mathbf{x}, \mathbf{v}) \in \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}^d. \quad (2.2.5)$$

The large crowd limits, which are assumed to exist, recover (2.2.1) with

$$\rho_\alpha(t, \mathbf{x}) := \lim_{N_\alpha \rightarrow \infty} \int_{\mathbb{R}^d} f_\alpha(t, \mathbf{x}, \mathbf{v}) d\mathbf{v}, \quad \rho_\alpha \mathbf{u}_\alpha(t, \mathbf{x}) := \lim_{N_\alpha \rightarrow \infty} \int_{\mathbb{R}^d} \mathbf{v} f_\alpha(t, \mathbf{x}, \mathbf{v}) d\mathbf{v}. \quad (2.2.6)$$

The pressure tensor $\mathbb{P}_\alpha$ corresponds to the second-order moments

$$\mathbb{P}_\alpha(t, \mathbf{x}) := \lim_{N_\alpha \rightarrow \infty} \int_{\mathbb{R}^d} (\mathbf{v} - \mathbf{u}_\alpha) \otimes (\mathbf{v} - \mathbf{u}_\alpha) f_\alpha(t, \mathbf{x}, \mathbf{v}) d\mathbf{v}, \quad (2.2.7)$$

Here, one encounters the familiar problem of lacking closure when dealing with the highest truncated $\mathbf{v}$ -moment of f_α . In the paper [58] the ‘pressure-less’ setup is pursued in the multi-species analysis. The two main aspect of this work is dealing with *arbitrary* pressure, without making any specifics about the second-order closure for $\mathbb{P}_\alpha$. We show that alignment enforces large time self-organization into flocking behavior while containing arbitrary internal fluctuations, along the one-species analysis in [141].

2.2.1.2 Connected arrays and alignment

As pointed out in [58], a key ingredient in the flocking behavior of multi-species is that their communication array, $\{\phi_{\alpha\beta}\}$, forms a *connected network*. To this end, we consider the symmetric radial array in (2.2.3), $K := \{k_{\alpha\beta}(r)\}$. Let $\Delta_{\mathcal{M}}K(r)$ be the *weighted graph Laplacian* associated with $K(r)$,

$$(\Delta_{\mathcal{M}}K(r))_{\alpha\beta} = \begin{cases} -k_{\alpha\beta}(r)\sqrt{M_{\alpha}M_{\beta}} & , \alpha \neq \beta \\ \sum_{\gamma \neq \alpha} k_{\alpha\gamma}(r)M_{\gamma} & , \alpha = \beta \end{cases} ,$$

where the weights, $\mathcal{M} = \{M_{\alpha}\}_{\alpha \in \mathcal{I}}$, consists of the masses of different species which remain constant in time

$$M_{\alpha} := \int \rho_{\alpha 0}(\mathbf{x}) d\mathbf{x} = \int \rho_{\alpha}(t, \mathbf{x}) d\mathbf{x}.$$

Ordering its real eigenvalues in an increasing order, one confirms that $\lambda_1(\Delta_{\mathcal{M}}K(r)) = 0$ corresponding to the constant eigenvector $\sqrt{\mathbf{M}} := \{\sqrt{M_{\alpha}}\}_{\alpha \in \mathcal{I}}$. The graph associated with $K(r)$ is *connected* as long as there is positive *spectral gap* associated with the weighted Laplacian $\lambda_2(\Delta_{\mathcal{M}}K(r)) > 0$. In particular, since $\Phi \succeq K$, and since the Fiedler number, [41, 43], $\lambda_2(\cdot)$, is an increasing function of its entries, the connectivity of K implies that Φ is connected.

Let $\rho_{\alpha}e_{\alpha}$ denote the internal energy of species α

$$\rho_{\alpha}e_{\alpha}(t, \mathbf{x}) := \int \frac{|\mathbf{v} - \mathbf{u}_{\alpha}|^2}{2} f_{\alpha}(t, \mathbf{x}, \mathbf{v}) d\mathbf{v} = \frac{1}{2} \text{trace}(\mathbb{P}_{\alpha}).$$

We investigate the flocking behavior of multi-species by studying the *energy fluctuations*,

$$\delta \mathcal{E}(t) = \sum_{\alpha, \beta \in \mathcal{I}} \iint \left(\frac{|\mathbf{u}_{\alpha}(t, \mathbf{x}) - \mathbf{u}_{\beta}(t, \mathbf{x}')|^2}{2} + e_{\alpha}(t, \mathbf{x}) + e_{\beta}(t, \mathbf{x}') \right) \rho_{\alpha}(t, \mathbf{x}) \rho_{\beta}(t, \mathbf{x}') d\mathbf{x}. \quad (2.2.8)$$

The first term quantifies the fluctuations of macroscopic velocities, $\mathbf{u}_{\alpha}(t, \cdot)$, the last two terms reflect the microscopic fluctuations, $|\mathbf{v} - \mathbf{u}_{\alpha}(t, \cdot)|^2$.

Our main result indicates the decay of the total energy fluctuations — both the macroscopic and internal (kinetic) fluctuations of the whole crowd, provided that the multi-species network is connected and there is a control on the spatial dispersion.

Theorem 2.2.1 *Let $(\rho_\alpha(t, \cdot), \mathbf{u}_\alpha(t, \cdot), \mathbb{P}_\alpha(t, \cdot)) \in (L^\infty \cap L^1_+(\mathbb{R}^d)) \times W^{1,\infty}(\mathbb{R}^d) \times W^{1,\infty}(\mathbb{R}^d)$, $\alpha \in \mathcal{I}$ be a strong solution to the multi-species dynamics (2.2.1)–(2.2.3), subject to compactly supported initial data $(\rho_{\alpha 0}, \mathbf{u}_{\alpha 0}, \mathbb{P}_{\alpha 0})$. Assume the communication array $\Phi = \{\phi_{\alpha\beta}\}$ has an entry-wise lower bound in terms of decreasing metric array, $\Phi(\mathbf{x}, \mathbf{x}') \succeq K(|\mathbf{x} - \mathbf{x}'|)$, with $k_{\alpha\beta}$ being decreasing, and the spectral gap of K admits a lower bound*

$$\lambda_2(\Delta_{\mathcal{M}} K(r)) \geq \eta(r). \quad (2.2.9)$$

Let $D(t)$ denote the spatial diameter,

$$D(t) := \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{S}(t)} |\mathbf{x} - \mathbf{x}'|, \quad \mathcal{S}(t) := \cup_{\alpha} \text{supp}\{\rho_{\alpha}(t, \cdot)\}.$$

Then there exists a constant, $C_M > 0$, such that the energy fluctuations satisfy the decay bound

$$\delta \mathcal{E}(t) \leq \exp \left\{ -2C_M \int_0^t \eta(D(s)) ds \right\} \delta \mathcal{E}(0). \quad (2.2.10)$$

We emphasize that Theorem 2.2.1 applies to general pressure tensors and any symmetric communication protocol satisfying (2.2.9): the kernels $\phi_{\alpha\beta}$ need not be metric nor bounded.

An immediate corollary is obtained under fat-tail connectivity.

Corollary 2.2.2 *Consider the multi-species system (2.2.1)–(2.2.3) associated with communication array $\Phi = \{\phi_{\alpha\beta}\}$ which admits a Pareto-type fat-tail bound*

$$\Phi(\mathbf{x}, \mathbf{x}') \succeq K(|\mathbf{x} - \mathbf{x}'|), \quad \lambda_2(\Delta_{\mathcal{M}} K(r)) \geq \frac{C_K}{(1+r)^\theta}, \quad \theta \leq 1. \quad (2.2.11)$$

Assume the crowd disperses at a rate of order $\gamma \geq 0$,

$$D(t) \leq C_D(1+t)^\gamma, \quad \gamma \geq 0, \quad (2.2.12)$$

If the fat-tail connectivity holds in the sense that $\gamma\theta \leq 1$, then there is a large time flocking behavior with decay rate

$$\delta\mathcal{E}(t) \leq \delta\mathcal{E}(0) \times \begin{cases} (1+t)^{-\nu_\theta}, & \gamma\theta = 1 \\ \exp\{-\nu_\theta t^{1-\gamma\theta}\}, & \gamma\theta < 1. \end{cases} \quad (2.2.13)$$

The decay constant ν_θ is specified with $\nu_\theta = 2C_M C_K C_D^{-\theta}$.

Notice that

$$\begin{aligned} \delta\mathcal{E}(t) &\geq \sum_{\alpha, \beta} \iint \frac{|\mathbf{u}_\alpha(t, \mathbf{x}) - \mathbf{u}_\beta(t, \mathbf{x}')|^2}{2} \rho_\alpha(t, \mathbf{x}) \rho_\beta(t, \mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= M \sum_\alpha \int |\mathbf{u}_\alpha(t, \mathbf{x}) - \mathbf{u}_\infty|^2 \rho_\alpha(t, \mathbf{x}) \, d\mathbf{x}, \quad \mathbf{u}_\infty := \frac{\sum_\alpha \int \rho_\alpha \mathbf{u}_\alpha}{M}, \end{aligned}$$

the bound (2.2.13) implies the flocking of different species towards the mean velocity $\mathbf{u}_\infty$,

$$\sum_{\alpha \in \mathcal{I}} \int |\mathbf{u}_\alpha(t, \mathbf{x}) - \mathbf{u}_\infty|^2 \rho_\alpha(t, \mathbf{x}) \, d\mathbf{x} \rightarrow 0, \quad t \rightarrow \infty.$$

2.2.2 Graph connectivity – weighted Poincaré inequalities

Given an $N \times N$ symmetric array $A = \{a_{\alpha\beta}\}$ consisting of non-negative entries, and the positive weights $\mathcal{W} = \{w_\alpha\}$. Let $\Delta_{\mathcal{W}} A$ denote the weighted Laplacian

$$(\Delta_{\mathcal{W}} A)_{\alpha\beta} = \begin{cases} -a_{\alpha\beta} \sqrt{w_\alpha w_\beta}, & \alpha \neq \beta \\ \sum_{\gamma \neq \alpha} a_{\alpha\gamma} w_\gamma, & \alpha = \beta \end{cases}. \quad (2.2.14)$$

We are concerned with an inequality of the form,

$$\sum_{\alpha, \beta} |\mathbf{x}_\alpha - \mathbf{x}_\beta|^2 w_\alpha w_\beta \geq \nu \sum_{\alpha, \beta} |\mathbf{x}_\alpha - \mathbf{x}_\beta|^2 w_\alpha w_\beta, \quad \nu > 0,$$

in order to quantify the effect of graph connectivity. To this end, we introduce weighted Poincaré inequalities without proofs. Details of derivations refer to [58].

Lemma 2.2.3 (Weighted Poincare Inequalities - vectors) *There holds*

$$\sum_{\alpha, \beta} a_{\alpha\beta} |\mathbf{x}_\alpha - \mathbf{x}_\beta|^2 w_\alpha w_\beta \geq \frac{\lambda_2(\Delta_{\mathcal{W}} A)}{\sum_{\beta} w_\beta} \sum_{\alpha, \beta} |\mathbf{x}_\alpha - \mathbf{x}_\beta|^2 w_\alpha w_\beta, \quad \mathcal{W} := \{w_\alpha\}. \quad (2.2.15)$$

Remark 2.2.4 Lemma 2.2.3 shows the optimality of weighted Laplacian, from which we obtain a Courant-Fisher-type characterization

$$\frac{\lambda_2(\Delta_{\mathcal{W}} A)}{\sum_{\beta} w_\beta} = \min_{|\delta \mathbf{x}|_w^2 = 1} \sum_{\alpha} \sum_{\beta \neq \alpha} |\mathbf{x}_\alpha - \mathbf{x}_\beta|^2 w_\alpha w_\beta, \quad |\delta \mathbf{x}|_w^2 := \sum_{\alpha} \sum_{\beta \neq \alpha} |\mathbf{x}_\alpha - \mathbf{x}_\beta|^2 w_\alpha w_\beta. \quad (2.2.16)$$

In the special case of equi-weighted species, $w_\alpha \equiv 1$, the weighted graph Laplacian reduces to the usual graph Laplacian, ΔA . Its second eigenvalue, $\lambda_2(\Delta A)$, known as the Fiedler number, quantifies the connectivity of the graph associated with adjacency matrix A [42]. If the number of species is finite, $|\mathcal{I}| < \infty$, there holds,

$$\frac{M}{\kappa^2 |\mathcal{I}|} \leq \frac{\lambda_2(\Delta_{\mathcal{W}} A)}{\lambda_2(\Delta A)} \leq \frac{M \kappa^2}{|\mathcal{I}|}, \quad \kappa = \frac{\max w_\alpha}{\min w_\alpha}, \quad M = \sum_{\alpha \in \mathcal{I}} w_\alpha,$$

hence the graph formed by A is connected if $\lambda_2(\Delta_{\mathcal{W}} A) \simeq \lambda_2(\Delta A) > 0$. The advantage of using the weighted graph Laplacian is that it produces the right scaling for the decay rate of fluctuations, which is independent of the condition number, κ , and the number of species, $|\mathcal{I}|$.

Inequality (2.2.15) yields a lower bound on the degree of each node. Fix an index γ and test the inequality (2.2.15) with the vector $\{\mathbf{x} | \mathbf{x}_\alpha = \begin{cases} 0 & \alpha \neq \gamma \\ \nu & \alpha = \gamma \end{cases}\}$, where normalization factor

$\nu = (2 \sum_{\beta \neq \gamma} w_\beta w_\gamma)^{-1/2}$ is taken so that $|\delta \mathbf{x}|_w = 1$, we obtain

$$\deg_\gamma(A) := \sum_{\gamma \neq \beta} a_{\gamma\beta} w_\beta \geq C_w \lambda_2(\Delta_{\mathcal{W}} A), \quad C_w = 1 - \frac{\max_\beta w_\beta}{\sum_\beta w_\beta} > 0. \quad (2.2.17)$$

Lemma 2.2.3 is extended from vectors to vector-functions as follows.

Lemma 2.2.5 (Weighted Poincare Inequalities - vector-functions) *Let $\{w_\gamma\}$ be non-negative weight functions with positive finite masses $M_\gamma := \int w_\gamma(\mathbf{x}) \, d\mathbf{x} > 0$. There holds*

$$\begin{aligned} \sum_{\alpha \neq \beta} a_{\alpha\beta} \iint |u_\alpha(\mathbf{x}) - u_\beta(\mathbf{x}')|^2 w_\alpha(\mathbf{x}) w_\beta(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ \geq \nu_A \sum_{\alpha, \beta} \iint |u_\alpha(\mathbf{x}) - u_\beta(\mathbf{x}')|^2 w_\alpha(\mathbf{x}) w_\beta(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}', \end{aligned} \quad (2.2.18)$$

with ν_A given by

$$\nu_A = \lambda_2(\Delta_{\mathcal{W}} A) \frac{C_M}{M}, \quad C_M = 1 - \frac{\max_\gamma M_\gamma}{M}, \quad M := \sum_\gamma M_\gamma.$$

It is noticed that the definition of weighted graph Laplacian (2.2.14) does not involve the diagonal terms of the adjacency matrix, $a_{\alpha\alpha}$. Hence the graph connectivity is independent of self-interactions with in the same specie.

2.2.3 Flocking of strong solutions

Now we study the decay of energy fluctuations in Theorem 2.2.1 and the subsequent flocking statement in Corollary 2.2.2. To this end, we introduce the balance of energy for multi-species (see Appendix B),

$$\begin{aligned} \partial_t(\rho_\alpha E_\alpha) + \nabla_{\mathbf{x}} \cdot (\rho_\alpha E_\alpha \mathbf{u}_\alpha + \mathbb{P}_\alpha \mathbf{u}_\alpha + \mathbf{q}_\alpha) = \\ - \sum_{\beta \in \mathcal{I}} \int \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (2E_\alpha(\mathbf{x}) - \mathbf{u}_\alpha(\mathbf{x}) \cdot \mathbf{u}_\beta(\mathbf{x}')) \rho_\alpha(\mathbf{x}) \rho_\beta(\mathbf{x}') \, d\mathbf{x}', \end{aligned} \quad (2.2.19)$$

with the heat flux $\mathbf{q}_\alpha := \int \frac{|\mathbf{v} - \mathbf{u}_\alpha|^2}{2} (\mathbf{v} - \mathbf{u}_\alpha) f_\alpha \, d\mathbf{v}$. The total energy of a specie α is given by the second moment

$$\rho_\alpha E_\alpha(t, \mathbf{x}) := \lim_{N_\alpha \rightarrow \infty} \int \frac{|\mathbf{v}|^2}{2} f_\alpha(\mathbf{x}, \mathbf{v}, t) \, d\mathbf{v}.$$

By writing $|\mathbf{v}|^2 = |\mathbf{v} - \mathbf{u}_\alpha|^2 + 2\mathbf{u}_\alpha \cdot (\mathbf{v} - \mathbf{u}_\alpha) + |\mathbf{u}_\alpha|^2$, the total energy is decomposed into the sum of the kinetic energy and the internal energy

$$\rho_\alpha E_\alpha = \frac{\rho_\alpha |\mathbf{u}_\alpha|^2}{2} + \rho_\alpha e_\alpha, \quad \rho_\alpha e_\alpha := \lim_{N_\alpha \rightarrow \infty} \frac{1}{2} \int |\mathbf{v} - \mathbf{u}_\alpha|^2 f_\alpha \, d\mathbf{v}.$$

The energy production on the right of (2.2.19) is driven by the enstrophy

$$- \sum_{\beta \in \mathcal{I}} \int \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (2E_\alpha(\mathbf{x}) - \mathbf{u}_\alpha(\mathbf{x}) \cdot \mathbf{u}_\beta(\mathbf{x}')) \rho_\alpha(\mathbf{x}) \rho_\beta(\mathbf{x}') \, d\mathbf{x}'.$$

Equation (2.2.19) implies the dissipation of the global energy, $\mathcal{E}(t) := \sum_{\alpha \in \mathcal{I}} \int \rho_\alpha E_\alpha \, d\mathbf{x}$.

Indeed, integrating (2.2.19) in $\mathbf{x}$ and summing over $\alpha \in \mathcal{I}$ yields

$$\begin{aligned} \frac{d}{dt} \mathcal{E}(t) &= - \sum_{\alpha, \beta \in \mathcal{I}} \iint \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (2E_\alpha(\mathbf{x}) - \mathbf{u}_\alpha(\mathbf{x}) \cdot \mathbf{u}_\beta(\mathbf{x}')) \rho_\alpha(\mathbf{x}) \rho_\beta(\mathbf{x}') \, d\mathbf{x} \, d\mathbf{x}' \\ &= - \sum_{\alpha, \beta \in \mathcal{I}} \iint \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (2E_\beta(\mathbf{x}') - \mathbf{u}_\beta(\mathbf{x}') \cdot \mathbf{u}_\alpha(\mathbf{x})) \rho_\beta(\mathbf{x}') \rho_\alpha(\mathbf{x}) \, d\mathbf{x} \, d\mathbf{x}' \\ &= - \sum_{\alpha, \beta \in \mathcal{I}} \iint \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (E_\alpha(\mathbf{x}) + E_\beta(\mathbf{x}') - \mathbf{u}_\alpha(\mathbf{x}) \cdot \mathbf{u}_\beta(\mathbf{x}')) \rho_\alpha(\mathbf{x}) \rho_\beta(\mathbf{x}') \, d\mathbf{x} \\ &= - \sum_{\alpha, \beta \in \mathcal{I}} \iint \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') \left(\frac{|\mathbf{u}_\alpha(\mathbf{x}) - \mathbf{u}_\beta(\mathbf{x}')|^2}{2} + e_\alpha(\mathbf{x}) + e_\beta(\mathbf{x}') \right) \rho_\alpha(\mathbf{x}) \rho_\beta(\mathbf{x}') \, d\mathbf{x} \leq 0. \end{aligned} \tag{2.2.20}$$

The third equality is obtained by interchanging $\mathbf{x}, \mathbf{x}'$ and α, β , where we also make use of the symmetry of communication kernels, $\phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') = \phi_{\alpha\beta}(\mathbf{x}', \mathbf{x}) = \phi_{\beta\alpha}(\mathbf{x}, \mathbf{x}')$.

Rearrange (2.2.8), the energy fluctuations can be written as

$$\delta \mathcal{E}(t) = 2M\mathcal{E}(t) - |\mathbf{m}_1|^2.$$

Recall that the total mass M and the total momentum $\mathbf{m}_1$ are constant in time, (2.2.20) yields the decay rate of fluctuations

$$\frac{d}{dt}\delta \mathcal{E}(t) = -2M \sum_{\alpha, \beta \in \mathcal{I}} \iint \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') \left(\frac{|\mathbf{u}_\alpha - \mathbf{u}'_\beta|^2}{2} + e_\alpha + e'_\beta \right) \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}'. \quad (2.2.21)$$

We start the proof of main theorems.

Proof of Theorem 2.2.1 Since $\Phi(\mathbf{x}, \mathbf{x}') \succeq K(|\mathbf{x} - \mathbf{x}'|)$, $k_{\alpha\beta}$ being radial decreasing, by (2.2.21) we have

$$\begin{aligned} \frac{d}{dt}\delta \mathcal{E}(t) &\leq -2M \sum_{\alpha, \beta \in \mathcal{I}} \iint k_{\alpha\beta}(|\mathbf{x} - \mathbf{x}'|) \left(\frac{|\mathbf{u}_\alpha - \mathbf{u}'_\beta|^2}{2} + e_\alpha + e'_\beta \right) \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}' \\ &\leq -2M \sum_{\alpha \neq \beta} k_{\alpha\beta}(D(t)) \iint \left(\frac{|\mathbf{u}_\alpha - \mathbf{u}'_\beta|^2}{2} + e_\alpha + e'_\beta \right) \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}' \\ &= -M \sum_{\alpha \neq \beta} k_{\alpha\beta}(D(t)) \iint |\mathbf{u}_\alpha - \mathbf{u}'_\beta| \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}' - 2M \sum_{\alpha} \sum_{\beta \neq \alpha} k_{\alpha\beta}(D(t)) M_\beta \int e_\alpha \rho_\alpha \, d\mathbf{x} \\ &\quad - 2M \sum_{\beta} \sum_{\alpha \neq \beta} k_{\alpha\beta}(D(t)) M_\alpha \int e'_\beta \rho'_\beta \, d\mathbf{x}' \\ &= -M \sum_{\alpha \neq \beta} k_{\alpha\beta}(D(t)) \iint |\mathbf{u}_\alpha - \mathbf{u}'_\beta| \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}' - 2M \sum_{\alpha} \deg_\alpha(K(D(t))) \int e_\alpha \rho_\alpha \, d\mathbf{x} \\ &\quad - 2M \sum_{\beta} \deg_\beta(K(D(t))) \int e'_\beta \rho'_\beta \, d\mathbf{x}' \\ &= I + II + III. \end{aligned} \quad (2.2.22)$$

The $\deg_\beta(K(D(t)))$ in the term III is obtained due to the symmetry $k_{\alpha\beta} = k_{\beta\alpha}$.

Lemma 2.2.5 leads to the bound for the first term

$$\begin{aligned}
I &\leq -M \cdot \lambda_2(\Delta_{\mathcal{M}}K(D(t))) \frac{C_M}{M} \sum_{\alpha,\beta} \iint |\mathbf{u}_\alpha - \mathbf{u}'_\beta|^2 \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}' \\
&\leq -2C_M \eta(D(t)) \sum_{\alpha,\beta} \iint \frac{|\mathbf{u}_\alpha - \mathbf{u}'_\beta|^2}{2} \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}',
\end{aligned} \tag{2.2.23}$$

where $\eta(r)$, by our assumption, is a lower bound of spectral gap $\lambda_2(\Delta_{\mathcal{M}}K(r))$. By the estimate of degree (2.2.17), the second term can be bounded as

$$\begin{aligned}
II &\leq -2M \cdot C_M \lambda_2(\Delta_{\mathcal{M}}K(D(t))) \sum_{\alpha} \int \rho_\alpha e_\alpha \, d\mathbf{x} \\
&\leq -2M \cdot C_M \eta(D(t)) \sum_{\alpha} \int \rho_\alpha e_\alpha \, d\mathbf{x} \\
&= -2C_M \eta(D(t)) \sum_{\alpha,\beta} \iint e_\alpha \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}'.
\end{aligned} \tag{2.2.24}$$

The third term can be bounded similarly

$$III \leq -2C_M \eta(D(t)) \sum_{\alpha,\beta} \iint e'_\beta \rho_\alpha \rho'_\beta \, d\mathbf{x} \, d\mathbf{x}'. \tag{2.2.25}$$

Substituting (2.2.23)–(2.2.25) into (2.2.22), we obtain

$$\frac{d}{dt} \delta \mathcal{E}(t) \leq -2C_M \eta(D(t)) \delta \mathcal{E}(t). \tag{2.2.26}$$

Applying Gronwall inequality yields the desired estimate (2.2.10). $\square$

Proof of Corollary 2.2.2 Take $\eta(r) = \frac{C_K}{(1+r)^\theta}$ in (2.2.10) and combine with (2.2.12), we have

$$\begin{aligned} \delta \mathcal{E}(t) &\leq \delta \mathcal{E}(0) \exp\left\{-2C_M C_K \int_0^t (1+D(s))^{-\theta} ds\right\} \\ &\leq \delta \mathcal{E}(0) \times \begin{cases} (1+t)^{-\nu_\theta}, & \gamma\theta = 1 \\ \exp\{-\nu_\theta t^{1-\gamma\theta}\}, & \gamma\theta < 1 \end{cases}, \end{aligned}$$

where $\nu_\theta = 2C_M C_K C_D^{-\theta}$. The desired estimate (2.2.13) is recovered. $\square$

An important aspect of the study is to investigate the expansion of the spatial diameter $D(t)$, of which the growth is bounded by uniform velocity fluctuation,

$$\frac{d}{dt} D(t) \leq \delta V(t), \quad \delta V(t) := \max_{\alpha, \beta \in \mathcal{I}} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{S}(t)} |\mathbf{u}_\alpha(t, \mathbf{x}) - \mathbf{u}_\beta(t, \mathbf{y})|. \quad (2.2.27)$$

Ideally, it is expected that the alignment is strong enough to keep the dynamics within a finite ball, $D(t) \leq D_+ < \infty$. The uniform boundedness of spatial diameter is proved for the case of mono-kinetic closure (see e.g. [57, 58]),

$$f_\alpha(t, \mathbf{x}, \mathbf{v}) \rightarrow \rho_\alpha \delta(\mathbf{v} - \mathbf{u}_\alpha), \quad N_\alpha \rightarrow \infty$$

which is realized by zero pressure, $\mathbb{P}_\alpha \equiv 0$. That is, (2.2.12) holds for $\gamma = 0$, leading to the exponential flocking rate, $\delta \mathcal{E}(t) \lesssim \delta \mathcal{E}(0) e^{-\nu_\theta t}$ for $\theta < 1$.

However, when a general pressure is involved, the uniform boundedness of diameter is not yet accessible. In practice, one may seek a weaker dispersion bound that allows a slow time growth. For instance, assuming the velocities are uniformly bounded,

$$\max_{\alpha \in \mathcal{I}} \sup_{\mathbf{x} \in \mathcal{S}(t)} |\mathbf{u}_\alpha(t, \mathbf{x})| \leq V_+ < \infty, \quad \forall t \geq 0,$$

we can bound the uniform fluctuation with $\delta V(t) \leq 2V_+$. Then the integration of (2.2.27) yields a linear growth, $D(t) \leq D_0 + 2V_+ t$. Although bounded velocities are expected for physical

modelling, the rigorous proof is still left open, except in *one-dimension*.

2.2.4 Boundedness of velocities

We particularly mention the flocking result on one-dimensional multi-species in the joint work with Eitan Tadmor [95]. A Sobolev bound of velocities is rigorously *proved* with respect to singular communications.

Let $\mathcal{S}_\alpha(t) = \text{supp}\{\rho_\alpha(t, \cdot)\}$ denote the support of the specie α , of which the diameter is represented by $D_\alpha(t) := \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{S}_\alpha(t)} |\mathbf{x} - \mathbf{x}'|$. Our discussion is based on the following assumptions:

(H1) The hydrodynamics (2.2.1) admits a strong solution governed by Neumann boundary conditions

$$\mathbf{u}_\alpha \cdot n_\alpha|_{\partial\mathcal{S}_\alpha} = 0, \quad \mathbb{P}_\alpha \cdot n_\alpha|_{\partial\mathcal{S}_\alpha} = 0, \quad \mathbf{q}_\alpha \cdot n_\alpha|_{\partial\mathcal{S}_\alpha} = 0, \quad \forall \alpha \in \mathcal{I}.$$

(H2) $\mathcal{S}_\alpha(t)$ has smooth boundary satisfying a Lipschitz or a cone condition, $\forall \alpha \in \mathcal{I}$.

(H3) The density of each specie inside their support is bounded away from vacuum,

$$\min_{\mathbf{x} \in \mathcal{S}_\alpha(t)} \rho_\alpha(t, \mathbf{x}) \geq \rho_- > 0.$$

We consider one-dimensional multi-species driven by singular self-interactions, $\phi_{\alpha\alpha}(r) = r^{-(1+2s)}$, $0 < s < 1$,

$$\left\{ \begin{array}{l} \partial_t \rho_\alpha + \partial_x(\rho_\alpha u_\alpha) = 0, \\ \partial_t(\rho_\alpha u_\alpha) + \partial_x(\rho_\alpha u_\alpha^2 + p_\alpha) = \\ \quad \int_{\mathcal{S}_\alpha(t)} \frac{(u_\alpha - u'_\alpha)}{|x' - x|^{1+2s}} \rho_\alpha \rho'_\alpha dx' + \sum_{\beta \neq \alpha} \int_{\mathcal{S}_\beta(t)} \phi_{\alpha\beta}(x, x')(u'_\beta - u_\alpha) \rho_\alpha \rho'_\beta dx' \end{array} \right. \quad (2.2.28)$$

Here the pressure tensor p_α and the heat flux q_α reduce to scalars.

Recall estimate (2.2.21) and ignore the negative contributions from internal energies and

cross interactions, we obtain

$$\frac{d}{dt}\delta\mathcal{E}(t) \leq -M \sum_{\alpha \in \mathcal{I}} \iint_{\mathcal{S}_\alpha(t) \times \mathcal{S}_\alpha(t)} \frac{|u_\alpha - u'_\alpha|^2}{|x - x'|^{1+2s}} \rho_\alpha \rho'_\alpha dx dx'. \quad (2.2.29)$$

This implies that velocities u_α are bounded in homogeneous Sobolev space $\dot{H}^s(\mathcal{S}_\alpha)$. Specifically, taking into account the non-vacuous assumption (H3), the integration of (2.2.29) yields

$$\begin{aligned} \int_0^t \sum_{\alpha} \|u_\alpha(\tau, \cdot)\|_{\dot{H}^s(\mathcal{S}_\alpha)}^2 d\tau &\leq C_\rho^2 \int_0^t \sum_{\alpha} \iint_{\mathcal{S}_\alpha(t) \times \mathcal{S}_\alpha(t)} \frac{|u_\alpha - u'_\alpha|^2}{|x - x'|^{1+2s}} \rho_\alpha \rho'_\alpha dx dx' d\tau \\ &\leq C_\rho^2 C_0, \quad C_\rho := \frac{1}{\rho_-}, \quad C_0 := \frac{\delta\mathcal{E}(0)}{M} \end{aligned} \quad (2.2.30)$$

Estimate (2.2.30) implies a dispersion bound sought in (2.2.12). Following the lines of [95], the dispersion of species is governed by

$$\frac{d}{dt}D_\alpha(t) \leq \delta V_\alpha(t), \quad \delta V_\alpha(t) := \sup_{x, x' \in \mathcal{S}_\alpha(t)} |u_\alpha(t, x) - u_\alpha(t, x')|.$$

Gagliardo-Nirenberg inequality [34] implies that for $\frac{1}{2} < s < 1$ there holds

$$|u_\alpha(t, x) - u_\alpha(t, x')| \leq C_s \|u_\alpha(t, \cdot)\|_{\dot{H}^s(\mathcal{S}_\alpha(t))} |x - x'|^{s-\frac{1}{2}}, \quad \forall x, x' \in \mathcal{S}_\alpha(t), \quad \frac{1}{2} < s < 1,$$

and hence $\frac{d}{dt}D_\alpha(t) \leq C_s \|u_\alpha(t, \cdot)\|_{\dot{H}^s(\mathcal{S}_\alpha(t))} D_\alpha(t)^{s-\frac{1}{2}}$. It follows that

$$\frac{d}{dt} \sum_{\alpha} D_\alpha^{\frac{3}{2}-s}(t) \leq C'_s \sum_{\alpha} \|u_\alpha(t, \cdot)\|_{\dot{H}^s(\mathcal{S}_\alpha(t))}, \quad C'_s = \left(\frac{3}{2} - s\right) C_s. \quad (2.2.31)$$

Since $\frac{3}{2} - s < 1$ then $D^{\frac{3}{2}-s}(t) \leq \left(\sum_{\alpha} D_\alpha(t)\right)^{\frac{3}{2}-s} \leq \sum_{\alpha} D_\alpha^{\frac{3}{2}-s}(t)$, and integration of (2.2.31)

yields

$$\begin{aligned}
D^{\frac{3}{2}-s}(t) &\leq \sum_{\alpha} D_{\alpha 0}^{\frac{3}{2}-s} + C'_s \int_0^t \sum_{\alpha} \|u_{\alpha}(\tau, \cdot)\|_{\dot{H}^s(\mathcal{S}_{\alpha}(t))} d\tau \\
&\leq \sum_{\alpha} D_{\alpha 0}^{\frac{3}{2}-s} + C'_s \left(\int_0^t \sum_{\alpha} \|u_{\alpha}(\tau, \cdot)\|_{\dot{H}^s(\mathcal{S}_{\alpha}(t))}^2 d\tau \right)^{\frac{1}{2}} \left(\int_0^t 1 d\tau \right)^{\frac{1}{2}} \\
&\leq \sum_{\alpha} D_{\alpha 0}^{\frac{3}{2}-s} + C'_s (C_{\rho}^2 C_0)^{\frac{1}{2}} t^{\frac{1}{2}}.
\end{aligned} \tag{2.2.32}$$

This concludes that the overall crowd is dispersed at a rate no faster than

$$D(t) \leq C_D (1+t)^{\gamma_s}, \quad \gamma_s = \frac{1}{3-2s}, \quad \frac{1}{2} < s < 1. \tag{2.2.33}$$

Assuming $\theta\gamma_s < 1$, the bound can be improved through a bootstrap argument (see Appendix C), which yields

$$D(t) \leq C_{D,\epsilon} (1+t)^{\gamma_{\epsilon}}, \quad \gamma_{\epsilon} = \frac{1+\epsilon}{3-2s+\theta}, \quad 0 < \epsilon \ll 1. \tag{2.2.34}$$

The flocking behavior is stated as follows.

Theorem 2.2.6 *Let $(\rho_{\alpha}, u_{\alpha}, \mathbb{P}_{\alpha})$ be a non-vacuous strong solution of (2.2.28) satisfying (H1)-(H3), with cross-interactions, $\Phi(x, x') \succeq K(|x - x'|)$. Assume that $K(r)$ admits a fat-tail*

$$\lambda_2(\Delta_{\mathcal{M}} K(r)) \geq \frac{C_K}{(1+r)^{\theta}}, \quad \theta \geq 0.$$

If the fat-tail condition holds in the sense that

$$\theta + 2s < 3, \quad \frac{1}{2} < s < 1, \tag{2.2.35}$$

then there is a large time flocking behavior with fractional exponential rate

$$\delta \mathcal{E}(t) \lesssim \delta \mathcal{E}(0) \exp\{-C_{\theta} t^{1-\theta\gamma_{\epsilon}}\}, \quad \gamma_{\epsilon} = \frac{1+\epsilon}{3-2s+\theta}, \quad C_{\theta} = 2C_M C_K (C_{D,\epsilon})^{-\theta},$$

where $\epsilon > 0$ is arbitrarily small.

Remark 2.2.7 *In the genuine multi-species set-up, the distinction between ‘singular’ self-interactions and ‘fat-tail’ cross-communications allows the respective controls on crowd expansion and network connectivity. However, as the situation degenerates to single-species, the ‘singular head’ and the ‘heavy tail’ are both imposed on the uniform communication kernel [141], i.e.*

$$\phi_{s,\theta}(x, x') \begin{cases} = |x - x'|^{-(1+2s)}, & |x - x'| \leq R, \quad 0 < s < 1, \\ \geq C_R(1 + |x - x'|)^{-\theta}, & |x - x'| > R. \end{cases}$$

Then the alignment behavior can still be obtained under condition (2.2.35).

Remark 2.2.8 *In general, if a function u is bounded in $W^{s,p}(\Omega)$, $\Omega \subseteq \mathbb{R}^d$,*

$$\|u\|_{W^{s,p}(\Omega)} := \|u\|_{L^p(\Omega)} + \left\{ \iint_{\Omega \times \Omega} \frac{|u(\mathbf{x}) - u(\mathbf{x}')|^p}{|\mathbf{x} - \mathbf{x}'|^{d+sp}} d\mathbf{x} d\mathbf{x}' \right\}^{\frac{1}{p}} < \infty,$$

Gagliardo-Nirenberg inequality implies Hölder continuity for $\frac{d}{p} < s < 1$,

$$|u(\mathbf{x}) - u(\mathbf{x}')| \leq C_s \|u_\alpha(t, \cdot)\|_{W^{s,p}(\Omega)} |\mathbf{x} - \mathbf{x}'|^{s - \frac{d}{p}}, \quad \forall \mathbf{x}, \mathbf{x}' \in \Omega, \quad \frac{d}{p} < s < 1.$$

In particular, the value $p = 2$ is taken in the velocity bound (2.2.30), hence the flocking statement is restricted to $d = 1$ -case.

In recent studies, the discussion of flocking dynamics is extended to p -alignment for single-species [142] and later on multi-species [95], where the Sobolev velocity bound can be extended to multi-dimensions. The p -alignment hydrodynamics is given by

$$\begin{cases} \partial_t \rho_\alpha + \nabla_{\mathbf{x}} \cdot (\rho_\alpha \mathbf{u}_\alpha) = 0, \\ \partial_t (\rho_\alpha \mathbf{u}_\alpha) + \nabla \cdot (\rho_\alpha \mathbf{u}_\alpha \otimes \mathbf{u}_\alpha + \mathbb{P}_\alpha) = \mathbf{A}_p(\rho, \mathbf{u}), \end{cases} \quad (2.2.36)$$

with p -alignment term

$$\begin{aligned} \mathbf{A}_p(\rho, \mathbf{u}) &:= \int \phi_{\alpha\alpha}(\mathbf{x}, \mathbf{x}') |\mathbf{u}'_\alpha - \mathbf{u}_\alpha|^{2p-2} (\mathbf{u}'_\alpha - \mathbf{u}_\alpha) \rho_\alpha \rho'_\alpha d\mathbf{x}' \\ &\quad + \sum_{\beta \neq \alpha} \int \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (\mathbf{u}'_\beta - \mathbf{u}_\alpha) \rho_\alpha \rho'_\beta d\mathbf{x}', \quad p \geq 1. \end{aligned}$$

The system (2.2.36) is driven by an entropic pressure: we say that the pressure $\mathbb{P}_\alpha$ is *entropic* if it has non-negative trace, $\rho_\alpha e_\alpha := \frac{1}{2} \text{trace}(\mathbb{P}_\alpha) \geq 0$, and satisfies

$$\partial_t(\rho_\alpha e_\alpha) + \nabla_{\mathbf{x}} \cdot (\rho_\alpha e_\alpha \mathbf{u}_\alpha + \mathbf{q}_\alpha) + \text{trace}(\mathbb{P}_\alpha \nabla \mathbf{u}_\alpha) \leq -\frac{1}{2} \int \phi_{\alpha\alpha}(\mathbf{x}, \mathbf{x}') ((2e_\alpha)^p + (2e'_\alpha)^p) \rho \rho' d\mathbf{x}'. \quad (2.2.37)$$

Here $\mathbf{q}_\alpha$ is an arbitrary C^1 -flux. The entropic description (2.2.37) is motivated by the second moment of mesoscopic distribution, it ensures the dissipation of total energy, $\rho_\alpha E_\alpha := \rho_\alpha e_\alpha + \frac{\rho_\alpha |\mathbf{u}_\alpha|^2}{2}$,

$$\begin{aligned} \frac{d}{dt} \sum_\alpha \int \rho_\alpha E_\alpha d\mathbf{x} &\leq -\frac{1}{2} \iint (\phi_{\alpha\alpha}(\mathbf{x}, \mathbf{x}') (|\mathbf{u}'_\alpha - \mathbf{u}_\alpha|^{2p} + (2e_\alpha)^p + (2e'_\alpha)^p) \rho_\alpha \rho'_\alpha d\mathbf{x} d\mathbf{x}' \\ &\quad - \sum_{\beta \neq \alpha} \int \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') \left(\frac{|\mathbf{u}'_\beta - \mathbf{u}_\alpha|^2}{2} + e_\alpha + e'_\beta \right) \rho_\alpha \rho'_\beta d\mathbf{x} d\mathbf{x}'. \end{aligned}$$

This includes the aforementioned pure alignment ($p = 1$) which admits an energy *equality* (2.2.19) associated with constitutively defined pressure (2.2.7). Notice that the *inequality* (2.2.37) is not concerned with details of specific thermodynamic closure, hence it applies to a large class of tensors beyond those realizable as second-order moments.

By imposing singular self-interactions, $\phi_{\alpha\alpha}(r) = r^{-(d+2sp)}$, $\frac{d}{2p} < s < 1$, one recovers $\dot{W}^{s,2p}$ -boundedness of velocities with respect to arbitrary entropic pressure, following the lines of (2.2.29) - (2.2.30),

$$\sum_{\alpha \in \mathcal{I}} \int_0^t \|\mathbf{u}(\tau, \cdot)\|_{\dot{W}^{s,2p}(\mathcal{S}_\alpha(t))}^{2p} d\tau \leq C_\rho^2 C_0.$$

The flocking statement is subsequently extended to multidimensional p -alignment of order $p >$

$\frac{d}{2}$. We refer the details to [\[95, 142\]](#).

Chapter 3: Simulation Methods for Collective Dynamics

In this chapter, we discuss the computation methods for collective dynamics. While a huge amount of self-organized systems have been proposed in numerous literature, the simulation techniques may be summarized under common frameworks, including particle (agent-based) simulations and hydrodynamics simulations. In particular, we consider the aforementioned Cucker-Smale dynamics as a concrete example.

Section 3.1 discusses the approach of particle simulations, where one keeps track of each particle by discretizing the agent-based model. We show that the forward Euler time stepping, as applied to the Cucker-Smale particle system, ensures the decay of fluctuations. Numerical examples are presented to visualize the flocking behavior.

Section 3.2 is devoted to the simulation of hydrodynamic models. The hydrodynamic limit of a particle system, if exists, is represented in the form of nonlinear transport equations. Classical numerical methods for conservation laws can be applied for non-oscillatory discretization. To this end, we give a brief review of the second order central schemes. The high order of accuracy of these schemes is achieved by employing piecewise linear MUSCL reconstruction, where appropriately constructed slope limiter plays a key role. In our work, we revisit the analytical properties of the smooth van Albada slope limiter, following the joint paper with Eitan Tadmor [147]. Its TVD stability is rigorously proved despite the lack of minmod structure, which is a common feature shared by most limiters of choice so far. Its notable improvement over the classical minmod limiter is verified through a series of numerical examples. The hydrodynamic alignment driven by Cucker-Smale dynamics is also demonstrated through computations.

3.1 Particle simulations of flocking dynamics

We consider the simulation of collective dynamics with particle schemes. This approach involves solving agent-based models given by a set of ordinary differential equations (ODEs) or stochastic differential equations (SDEs). Classical ODE/SDE solvers can be applied to discretize the time-continuous system. Here we focus on the Cucker-Smale dynamics (2.1.1) as a representative example.

It is important to note that the emergence of flocking is much more dependent on the communication kernel than on the accuracy of the numerical scheme. Thus it suffices to apply the simple forward Euler method to simulate the large time behavior. The evolution of discrete solutions, $(\mathbf{x}_i^n, \mathbf{v}_i^n)$, of the system (2.1.1) generated with forward Euler time stepping is described by

$$\begin{cases} \mathbf{x}_i^{n+1} = \mathbf{x}_i^n + \Delta t \mathbf{v}_i^n \\ \mathbf{v}_i^{n+1} = \mathbf{v}_i^n + \frac{\Delta t}{N} \sum_{j=1}^N \phi_{ij}^n \cdot (\mathbf{v}_j^n - \mathbf{v}_i^n) \end{cases}, \quad \phi_{ij}^n := \phi(\mathbf{x}_i^n, \mathbf{x}_j^n), \quad 1 \leq i \leq N, \quad (3.1.1)$$

subject to initial data, $(\mathbf{v}_i^0, \mathbf{x}_i^0)$. When a symmetric kernel, $\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{y}, \mathbf{x})$, is considered, summing the second equation of (3.1.1) over $1 \leq i \leq N$ yields the conservation of momentum,

$$\sum_{i=1}^N \mathbf{v}_i^{n+1} = \sum_{i=1}^N \mathbf{v}_i^n.$$

We denote the constant mean velocity by $\bar{\mathbf{v}} := \sum_{i=1}^N \mathbf{v}_i^0 / N$.

Remark 3.1.1 *In contrast to the time-continuous model (2.1.1), the system (3.1.1) describes the occasional response of each agent to the influence from others, namely, the velocities are updated with a period of Δt instead of continuously in time. The discrete dynamics may serve as a better modelling for animal creatures, where the ‘delaying’ mechanism mimics the reaction time of individuals.*

We prove the decay of the fluctuations.

Theorem 3.1.2 Assume $(\mathbf{x}_i^n, \mathbf{v}_i^n)$, $1 \leq i \leq N$, is a solution to the discrete particle system (3.1.1) associated with a symmetric and uniformly bounded communication kernel, $0 < \phi_- \leq \phi \leq \phi_+$. Then the velocity fluctuation,

$$\delta V^n := \sum_{1 \leq i, j \leq N} |\mathbf{v}_i^n - \mathbf{v}_j^n|^2, \quad (3.1.2)$$

satisfies the estimate,

$$\delta V^{n+1} \leq (1 - 2\phi_- \Delta t + 2\phi_+^2 \Delta t^2) \cdot \delta V^n. \quad (3.1.3)$$

Proof. The velocity fluctuations can be written as

$$\delta V^n = 2N \sum_{i=1}^N |\mathbf{v}_i^n|^2 - 2N^2 |\bar{\mathbf{v}}|^2,$$

where the second term, due to the conservation of momentum, is a constant. Denote $\delta \mathbf{v}_i^n := \mathbf{v}_i^{n+1} - \mathbf{v}_i^n$, by the second equation of (3.1.1),

$$\begin{aligned} \delta V^{n+1} - \delta V^n &= 2N \sum_{i=1}^N (|\mathbf{v}_i^{n+1}|^2 - |\mathbf{v}_i^n|^2) \\ &= 2N \sum_{i=1}^N (2\mathbf{v}_i^n \cdot \delta \mathbf{v}_i^n + |\delta \mathbf{v}_i^n|^2) \\ &= 4\Delta t \sum_{1 \leq i, j \leq N} \phi_{ij}^n \mathbf{v}_i^n \cdot (\mathbf{v}_j^n - \mathbf{v}_i^n) + \frac{2\Delta t^2}{N} \sum_{i=1}^N \left| \sum_{j=1}^N \phi_{ij}^n (\mathbf{v}_j^n - \mathbf{v}_i^n) \right|^2 \\ &= I + II \end{aligned}$$

By the symmetry of ϕ , the first term on the right is bounded with

$$\begin{aligned}
I &= 4\Delta t \sum_{1 \leq i, j \leq N} \phi_{ij}^n \mathbf{v}_i^n \cdot (\mathbf{v}_j^n - \mathbf{v}_i^n) \\
&= 4\Delta t \sum_{1 \leq i, j \leq N} \phi_{ij}^n \mathbf{v}_j^n \cdot (\mathbf{v}_i^n - \mathbf{v}_j^n) \\
&= -2\Delta t \sum_{1 \leq i, j \leq N} \phi_{ij}^n |\mathbf{v}_i^n - \mathbf{v}_j^n|^2 \\
&\leq -2\Delta t \phi_- \sum_{1 \leq i, j \leq N} |\mathbf{v}_i^n - \mathbf{v}_j^n|^2 \\
&= -2\Delta t \phi_- \cdot \delta V^n.
\end{aligned}$$

We use the Cauchy-Schwartz's inequality to bound the second term,

$$\begin{aligned}
II &= \frac{2\Delta t^2}{N} \sum_{i=1}^N \left| \sum_{j=1}^N \phi_{ij}^n (\mathbf{v}_j^n - \mathbf{v}_i^n) \right|^2 \\
&\leq \frac{2\Delta t^2 \phi_+^2}{N} \sum_{i=1}^N \left(\sum_{j=1}^N |\mathbf{v}_j^n - \mathbf{v}_i^n| \right)^2 \\
&\leq \frac{2\Delta t^2 \phi_+^2}{N} \sum_{i=1}^N \left(N \sum_{j=1}^N |\mathbf{v}_i^n - \mathbf{v}_j^n|^2 \right) \\
&= 2\Delta t^2 \phi_+^2 \cdot \delta V^n
\end{aligned}$$

Combine the above estimates,

$$\begin{aligned}
\delta V^{n+1} - \delta V^n &\leq -2\Delta t \phi_- \cdot \delta V^n + 2\Delta t^2 \phi_+^2 \cdot \delta V^n, \\
\delta V^{n+1} &\leq (1 - 2\phi_- \Delta t + 2\phi_+^2 \Delta t^2) \cdot \delta V^n.
\end{aligned}$$

The positivity of ϕ_- ensures that the prefactor on the right is smaller than 1, at least when the time step Δt is sufficiently small. Hence the decay of fluctuations is recovered. $\square$

Remark 3.1.3 The velocity fluctuations (3.1.2) can be written as

$$\delta V^n = 2N \sum_{i=1}^N |\mathbf{v}_i^n - \bar{\mathbf{v}}|^2, \quad \bar{\mathbf{v}} := \frac{\sum_{i=1}^N \mathbf{v}_i^0}{N}.$$

Hence it follows from estimate (3.1.3) that the velocities of agents are aligned with the mean velocity $\bar{\mathbf{v}}$ asymptotically in time.

Remark 3.1.4 In Theorem 3.1.2 we assumed the strong uniform boundedness on ϕ . In general, for locally bounded communication kernels, $0 < \phi_-^n \leq \phi_{ij}^n \leq \phi_+^n$, the estimate (3.1.3) would be modified,

$$\begin{aligned} \delta V^n &\leq (1 - 2\phi_-^{n-1}\Delta t + 2(\phi_+^{n-1})^2\Delta t^2)\delta V^{n-1} \\ &\leq \prod_{k=0}^{n-1} (1 - 2\phi_-^k\Delta t + 2(\phi_+^k)^2\Delta t^2)\delta V^0 \\ &\lesssim \delta V^0 \exp\left\{-2 \sum_{k=0}^{n-1} \phi_-^k\Delta t\right\}. \end{aligned}$$

The last inequality is obtained by neglecting the high order terms, $O(\Delta t^2)$, and using the inequality $1 + x \leq e^x$. Notice that the bound is a discrete analogue of the time-continuous counterpart (2.1.6). Hence we can expect the forward Euler method to recover similar flocking rates with those in Theorem 2.1.3.

We point out that following the lines of Theorem 3.1.2, one can also prove the decay of discrete velocity fluctuations for multi-species,

$$\delta V^n = \sum_{\alpha, \beta \in \mathcal{I}} \frac{1}{N_\alpha} \frac{1}{N_\beta} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} |\mathbf{v}_\alpha^{i,n} - \mathbf{v}_\beta^{j,n}|^2,$$

where $\mathbf{v}_\alpha^{i,n}$ are generated with the forward Euler discretization of system (2.2.4).

Now we apply the scheme (3.1.1) to verify the flocking results for single-species, as proved in Theorem 2.1.3. Here we consider a two-dimensional initial configuration where $N = 20$

agents are randomly and uniformly generated within the rectangular domain, $[-3, 3]^2$. The initial velocity of each agent is a 2D vector, with each component generated independently according to the uniform distribution on $[0, 5]$. We examine the collective dynamics driven by the kernel

$$\phi(r) = \frac{1}{(1+r)^\theta}, \quad \theta > 0.$$

We simulate the motions of particles under the time step, $\Delta t = 0.1$, up to the terminal time, $T = 5$.

Figure 3.1.2 compares the trajectories of particles under different values of θ . It is observed that the motions of particles are gradually aligned with the trajectory of mass center for $\theta \leq 1$, as expected by the theory. When $\theta > 1$, however, the trend of alignment is much less obvious due to the weaker communications. The histories of velocity fluctuations are shown in Figure 3.1.1. An exponential decay in δV is observed in the case of $\theta = 0.5$ and a polynomial decay is recovered when $\theta = 1.0$. The statements of Theorem 2.1.3 are verified. In the case $\theta = 1.5$, which is beyond the existing discussions, the velocity fluctuation is decaying much slower and there is no clear tendency of flocking, $\delta V \rightarrow 0$. Indeed, the emergence of flocking for $\theta > 1$ is dependent on the initial configuration.

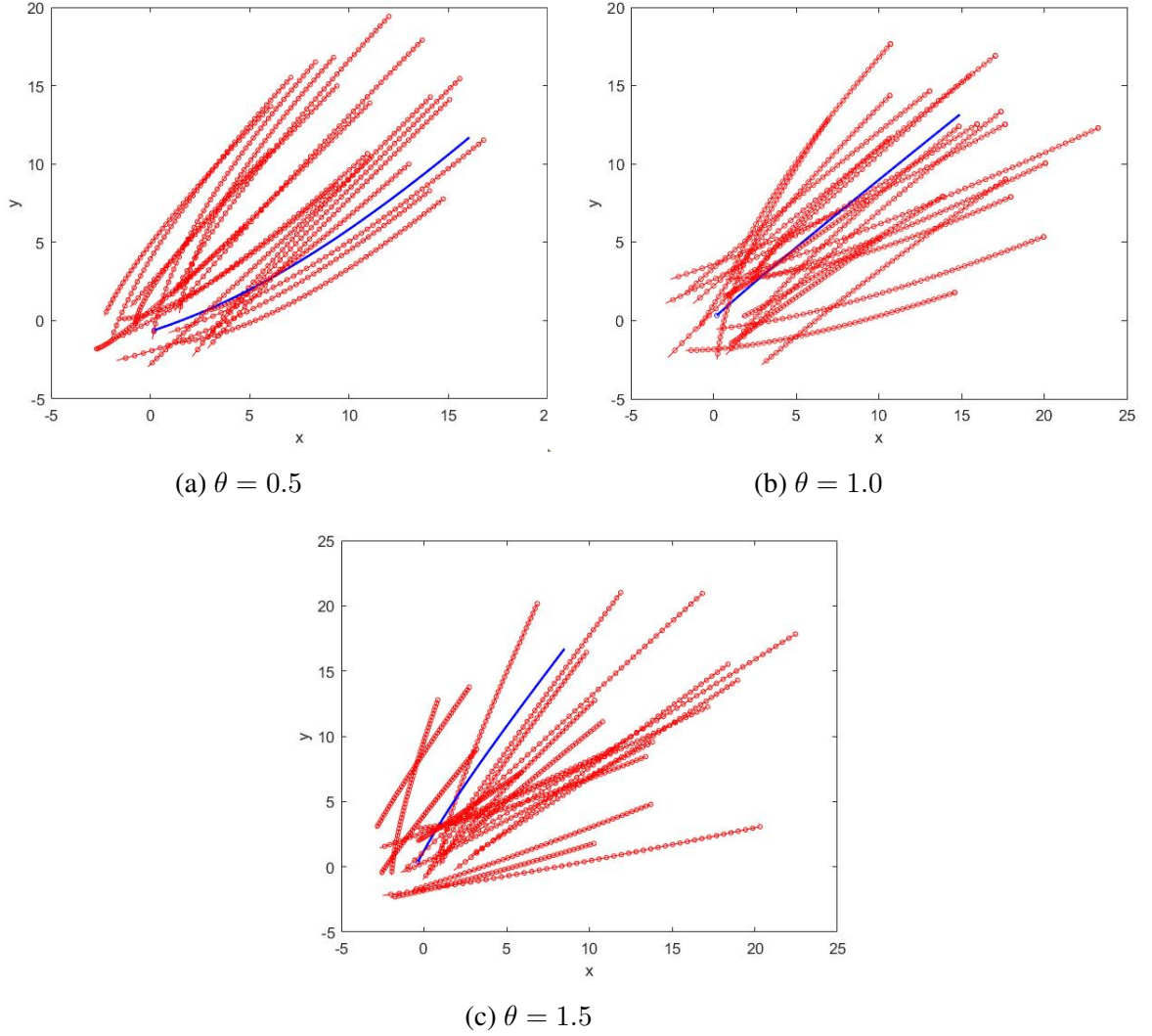
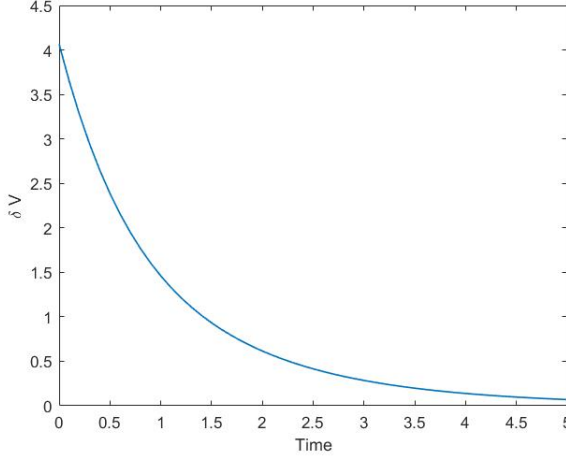


Figure 3.1.1: 2D Cucker-Smale, trajectories of $N = 20$ particles up to $T = 5$. Red lines: trajectories of particles; blue lines: trajectories of mass centers.

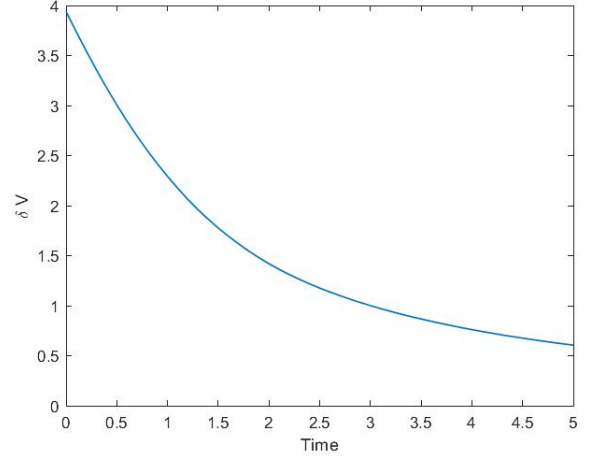
3.2 Hydrodynamic simulations – numerical methods for conservation laws

Although the approach of particle simulations has advantage in its simplicity, when a huge number of particles are involved it can be very expensive to simulate the evolution of each single individual. Alternatively, it would be more economical (in low-dimensions) to simulate large-crowd behaviors by discretizing the hydrodynamic models, in which case the *average* behavior of large population is emphasized.

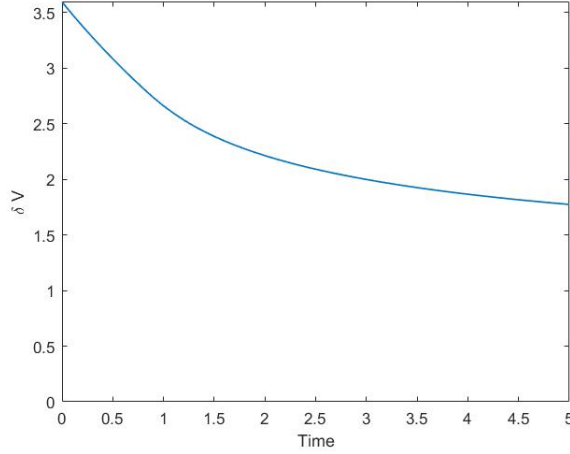
The collective hydrodynamics are governed by nonlinear transport equations. For instance,



(a) $\theta = 0.5$



(b) $\theta = 1.0$



(c) $\theta = 1.5$

Figure 3.1.2: 2D Cucker-Smale, history of velocity fluctuations. x -axis: computation time; y -axis: velocity fluctuations.

the single- and multi-species hydrodynamics, (2.1.14) and (2.2.1), are given by Euler systems associated with alignment forcing. The nonlinear advection terms must be appropriately discretized to prevent numerical oscillations near potential discontinuities. To this end, we discuss the numerical methods for conservation laws taking the form

$$\frac{\partial}{\partial t} \mathbf{u}(t, \mathbf{x}) + \nabla \cdot \mathbf{f}(\mathbf{u}(t, \mathbf{x})) = 0, \quad \mathbf{x} \in \Omega \subseteq \mathbb{R}^d, \quad t \geq 0. \quad (3.2.1)$$

Here, $\mathbf{u}(t, \mathbf{x}) = [u_1(t, \mathbf{x}), \dots, u_k(t, \mathbf{x})]^\top \in \mathbb{R}^k$ is the scalar/vector of conservative variables,

$\mathbf{f} = [f^1, \dots, f^d] \in \mathbb{R}^{k \times d}$ is the flux function. The inclusion of source terms would then be straightforward.

It is well-known that the solutions of nonlinear conservation laws can develop discontinuities even if the initial data is smooth. To prevent the numerical oscillations near discontinuities and extrema, there have been a lot of researches on non-oscillatory numerical schemes. The early stage frameworks include the monotone schemes [54] and the E-schemes [111], which inherit the monotonicity and the entropy stability of entropy solutions to scalar conservation laws. To break the barrier of first order accuracy, less restrictive monotonicity conditions are usually considered, the most popular among which is the *Total Variation Diminishing* (TVD) [52].

Suppose we have one-dimensional approximate solution to the scalar problem (3.2.1) under mesh size Δx and at time level $t^n = n\Delta t$,

$$u^n = \sum_j u_j^n \mathbf{1}_{[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]}, \quad x_{j \pm \frac{1}{2}} = x_j \pm \frac{\Delta x}{2},$$

where u_j^n is the approximation to the point value $u(t^n, x_j)$ /cell average

$$\bar{u}(t^n, x_j) := \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t^n, x) dx,$$

in the finite difference/finite volume framework. The solution u^n is said to be TVD if

$$\|u^n\|_{TV} := \sum_j |\Delta u_{j+\frac{1}{2}}^n| \leq \|u^{n-1}\|_{TV}, \quad \Delta u_{j+\frac{1}{2}}^n := u_{j+1}^n - u_j^n. \quad (3.2.2)$$

The TVD criteria has profound influence on the development of high resolution numerical schemes, including the second order flux-limiter schemes [137], MUSCL schemes [149], as well as even higher order TVB (*Total Variation Bounded*) methods, such as the UNO scheme [53], the MP scheme [136] and the nowadays popular ENO/WENO type schemes [123], etc.

In the real world applications such as computational fluid dynamics, the second order MUSCL schemes are widely applied due to its high resolution, efficiency and flexibility to the

mesh structure. The key ingredient of MUSCL schemes is to obtain a non-oscillatory piecewise linear polynomial reconstruction, $u^{\Delta x}(t^n, x) \approx u(t^n, x)$, from the cell averages $\{\bar{u}_j^n\}$,

$$\begin{aligned} u^{\Delta x}(t^n, x) &= \sum_j p_j^n(x) \mathbf{1}_{[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]}, \\ p_j^n(x) &= \bar{u}_j^n + \frac{u'_j}{\Delta x} (x - x_j). \end{aligned} \quad (3.2.3)$$

Here, u'_j is an approximate slope that satisfies the condition

$$\frac{u'_j}{\Delta x} = \frac{\partial}{\partial x} u(t^n, x_j) + O(\Delta x), \quad (3.2.4)$$

in which case we obtain a second order of accuracy in space. In general, the approximate slope u'_j is a function of the one-sided differences, which takes the form

$$u'_j = \psi(\Delta \bar{u}_{j-\frac{1}{2}}^n, \Delta \bar{u}_{j+\frac{1}{2}}^n).$$

The function ψ is known as the *slope limiter*. A variety of TVD limiters have been proposed and studied since 1980s' (see e.g. [22, 62, 63, 69, 98, 147]), well-known examples include the Min-Mod (MM) slope limiter

$$\psi_{MM}(a, b) = \text{minmod}(a, b) := \frac{\text{sgn}(a) + \text{sgn}(b)}{2} \cdot \min\{|a|, |b|\}, \quad (3.2.5)$$

and the one-parameter family of Generalized Min-Mod (GMM) limiter,

$$\psi_{GMM}(a, b) = \text{minmod}(\theta a, \frac{a+b}{2}, \theta b), \quad 1 \leq \theta \leq 2. \quad (3.2.6)$$

Once the piecewise linear interpolant (3.2.3) is obtained, it will be combined with a proper finite volume scheme to generate high resolution numerical solutions.

3.2.1 Finite volume methods: upwind vs central

Among finite volume methods there are two main classes: the *upwind* schemes and the *central* schemes. The main idea of upwind schemes is to employ upstream-biased information to approximate spatial derivative, whereas the central schemes process the information from upstream and downstream in an average manner.

The prototype of upwind schemes is the Godunov scheme [48]. The method computes a piecewise constant approximate solution to problem (3.2.1), $u^{\Delta x}(t, x)$, over cells of width $\Delta x = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, which takes the form

$$u^{\Delta x}(t^n, x) = \bar{u}_j^n, \quad x_{j-\frac{1}{2}} < x < x_{j+\frac{1}{2}}.$$

Here $\bar{u}_j^n$ is the approximation to the cell average,

$$\bar{u}_j^n \approx \bar{u}(t^n, x_j) := \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t^n, x) dx.$$

The calculation of $\bar{u}_j^{n+1}$ is based on the integration of (3.2.1) over the rectangle $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times [t^n, t^{n+1}]$, which yields

$$\begin{aligned} \bar{u}_j^{n+1} &= \bar{u}_j^n - \frac{1}{\Delta x} \int_{t^n}^{t^{n+1}} [f(u^{\Delta x}(t, x_{j+\frac{1}{2}})) - f(u^{\Delta x}(t, x_{j-\frac{1}{2}}))] dt, \\ &= \bar{u}_j^n - \lambda (f(u_{j+\frac{1}{2}}^*) - f(u_{j-\frac{1}{2}}^*)) \end{aligned} \quad (3.2.7)$$

where $\lambda := \frac{\Delta t}{\Delta x}$ is the fixed mesh ratio. The states,

$$u_{j+\frac{1}{2}}^* = u_{j+\frac{1}{2}}^*(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+), \quad u_{j-\frac{1}{2}}^* = u_{j-\frac{1}{2}}^*(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+),$$

are obtained with exact or approximate Riemann solvers (see e.g. [37, 39, 55, 120, 145]). Here, $u_{j+\frac{1}{2}}^-$ and $u_{j+\frac{1}{2}}^+$ are the reconstructed variables at $x_{j+\frac{1}{2}}$ from the left and the right. For systems

of equations, both left- and right-running waves may exist at interfaces. Hence the local characteristic decomposition of the flux $\mathbf{f}$ is required for the upwind constructions, based on which the ‘direction of the wind’ is identified field by field.

In the context of collective dynamics, the characteristic decomposition may encounter the problem of degeneration. For example, consider the one-dimensional single-species hydrodynamics (2.1.14) driven by vanishing pressure. The Jacobian of the flux function, $\mathbf{f} = \begin{bmatrix} \rho u \\ \rho u^2 \end{bmatrix}$, is given by

$$\frac{\partial \mathbf{f}}{\partial \mathbf{U}} = \begin{bmatrix} 0 & 1 \\ -u^2 & 2u \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \rho \\ \rho u \end{bmatrix},$$

associated with repeated eigenvalues, λ_i , and eigenvectors, $\mathbf{r}_i$,

$$\lambda_1 = \lambda_2 = u, \quad \mathbf{r}_1 = \mathbf{r}_2 = \begin{bmatrix} 1 \\ u \end{bmatrix}.$$

The singularity of the eigenvector matrix, $\mathbf{R} := [\mathbf{r}_1 \mathbf{r}_2] = \begin{bmatrix} 1 & 1 \\ u & u \end{bmatrix}$, invalidates the definition of characteristic variables, $\mathbf{R}^{-1}\mathbf{U}$, hence the characteristic-wise upwind constructions can not be applied.

To avoid the complex and sometimes impractical characteristic decompositions in general applications, one may take the alternative approach to calculate cell averages, i.e. the central schemes. The forerunner of central schemes is the celebrated Lax-Friedrich (LxF) scheme [89]:

$$\bar{u}_{j+\frac{1}{2}}^{n+1} = \frac{\bar{u}_j^n + \bar{u}_{j+1}^n}{2} - \lambda(f(u_{j+1}^n) - f(u_j^n)). \quad (3.2.8)$$

In contrast to the Godunov scheme (3.2.7), the LxF solver computes the cell averages on the staggered mesh, $\bar{u}(t, x_{j+\frac{1}{2}}) := \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u(t, x) dx$. The integration of equation (3.2.1) over the

staggered volume $[t^n, t^{n+1}] \times [x_j, x_{j+1}]$ yields

$$\bar{u}(t^{n+1}, x_{j+\frac{1}{2}}) = \frac{1}{\Delta x} \int_{x_j}^{x_{j+1}} u(t^n, x) dx - \frac{1}{\Delta x} \int_{t^n}^{t^{n+1}} [f(u(t, x_{j+1})) - f(u(t, x_j))] dt. \quad (3.2.9)$$

It is clear that the LxF scheme is a first order approximation to (3.2.9). Compared with the upwind framework, the advantage of central differencing is that the solution is smooth in the neighborhood of points $\{x_j\}$. Indeed, under a sufficiently small time step, Δt , the Riemann waves initiating from $x_{j+\frac{1}{2}}$ do not affect the adjacent cell centers within the time slab, $t \in [t^n, t^{n+1}]$. In this way, one avoids solving Riemann problems at x_j and hence the costly characteristic decompositions are not necessary.

Because of the simplicity and flexibility, central schemes would be suggested for applications to collective dynamics.

3.2.2 A brief review of second order central schemes

In spite of the great simplicity, the main disadvantage of the LxF scheme is the excessive numerical dissipation which damages the resolution at discontinuities and smooth extrema. The situation can be improved by applying higher order piecewise polynomial interpolations. We briefly review the second order fully-discrete and semi-discrete central schemes.

3.2.2.1 The Nessyahu-Tadmor scheme

The second order fully-discrete central scheme was first proposed by Nessyahu and Tadmor in 1990 [107]. The second order of accuracy is obtained at two aspects. Firstly, the solution $u(t^n, x)$ is approximated with piecewise linear MUSCL interpolant as given in (3.2.3). Integrating the equation (3.2.1) over the rectangle $[t^n, t^{n+1}] \times [x_j, x_{j+1}]$ yields

$$\bar{u}_{j+\frac{1}{2}}^{n+1} = \frac{1}{\Delta x} \left[\int_{x_j}^{x_{j+\frac{1}{2}}} p_j^n(x) dx + \int_{x_{j+\frac{1}{2}}}^{x_{j+1}} p_{j+1}^n(x) dx \right] - \frac{1}{\Delta x} \int_{t^n}^{t^{n+1}} [f(u^{\Delta x}(t, x_{j+1})) - f(u^{\Delta x}(t, x_j))] dt.$$

Secondly, the time integral of flux function can be approximated by the midpoint rule at the expense of $O(\Delta t^3)$ local truncation error. This results in the *NT scheme*

$$\bar{u}_{j+\frac{1}{2}}^{n+1} = \frac{\bar{u}_j^n + \bar{u}_{j+1}^n}{2} + \frac{1}{8}(u'_j - u'_{j+1}) - \lambda[f(u_{j+1}^{n+\frac{1}{2}}) - f(u_j^{n+\frac{1}{2}})], \quad (3.2.10)$$

The midpoint values, $u_j^{n+\frac{1}{2}}$, are predicted by Taylor expansions

$$u_j^{n+\frac{1}{2}} = \bar{u}_j^n - \frac{\lambda}{2}f'_j,$$

where the numerical derivative f'_j satisfies the second order accuracy condition

$$\frac{f'_j}{\Delta x} = \frac{\partial f}{\partial x}(\bar{u}_j^n) + O(\Delta x). \quad (3.2.11)$$

Different options of evaluating f'_j were proposed in [107]. For example, the numerical derivative can be computed in terms of exact flux Jacobian,

$$f'_j = A(\bar{u}_j^n)u'_j, \quad A(u) := \frac{\partial f}{\partial u}(u). \quad (3.2.12)$$

Alternatively, one can apply Jacobian-free approximations when the characteristic decomposition is computationally expensive (even sometimes inaccessible), e.g.

$$f'_j = \text{minmod}(\Delta f_{j-\frac{1}{2}}, \Delta f_{j+\frac{1}{2}}), \quad \Delta f_{j+\frac{1}{2}} := f(\bar{u}_{j+1}^n) - f(\bar{u}_j^n). \quad (3.2.13)$$

Numerical results in [71, 93] have shown that the Jacobian-free version of the central scheme does not deteriorate the high resolution. When applied to systems of equations, the NT scheme inherits the simplicity of the LxF solver, i.e. the constructions can be extended in a component-wise manner without the use of Riemann solvers or characteristic decompositions.

The multidimensional formulation of the NT scheme was obtained with similar integration procedures, details can be found in [71]. The order of accuracy can be further improved by

applying higher order piecewise polynomial interpolants, see e.g. [90, 91, 92, 93].

3.2.2.2 The semi-discrete formulation

It can be shown that the numerical dissipation of the NT scheme has order $O(\Delta x^4/\Delta t)$. In the convective problems where $\Delta t \sim \Delta x$, the NT scheme achieves higher resolution than the first order LxF scheme due to the reduced numerical viscosity. However, it is noticed that the NT scheme and its higher order extensions do not admit any semi-discrete limits, hence they are not appropriate for small time step computations or steady-state calculations. This motivated the development of semi-discrete central schemes. We sketch the derivations along the lines of [86, 87].

We start with a piecewise polynomial approximation $u^{\Delta x}(t^n, x) \approx u(t^n, x)$ of the form

$$u^{\Delta x}(t^n, x) = \sum_j p_j^n(x) \mathbf{1}_{[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]}.$$

The polynomials $p_j^n(x)$ should have desired order of accuracy and conserve the cell averages $\bar{u}_j^n := \bar{u}(x_j, t^n)$, i.e.

$$\frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p_j^n(x) dx = \bar{u}_j^n.$$

We denote the reconstructed variables at $x_{j+\frac{1}{2}}$ from the left and the right by

$$u_{j+\frac{1}{2}}^- := p_j^n(x_{j+\frac{1}{2}}), \quad u_{j+\frac{1}{2}}^+ := p_{j+1}^n(x_{j+\frac{1}{2}}). \quad (3.2.14)$$

Particularly, when the MUSCL interpolant (3.2.3) is applied we have

$$u_{j+\frac{1}{2}}^- := \bar{u}_j^n + \frac{1}{2} u'_j, \quad u_{j+\frac{1}{2}}^+ := \bar{u}_{j+1}^n - \frac{1}{2} u'_{j+1}. \quad (3.2.15)$$

To obtain a semi-discrete formulation the key idea is to take cell averaging over narrower control volumes. Assume that the maximal forward and backward wave speeds at $x_{j+\frac{1}{2}}$ are

estimated by $a_{j+\frac{1}{2}}^+ \geq 0$ and $a_{j+\frac{1}{2}}^- \leq 0$. At the next time level $t^{n+1} := t^n + \Delta t$, the region influenced by the Riemann fan originating at $x_{j+\frac{1}{2}}$ is approximated with the interval, $I_{j+\frac{1}{2}} := [x_{j+\frac{1}{2}}^-, x_{j+\frac{1}{2}}^+]$, where

$$x_{j+\frac{1}{2}}^+ := x_{j+\frac{1}{2}} + a_{j+\frac{1}{2}}^+ \Delta t, \quad x_{j+\frac{1}{2}}^- := x_{j+\frac{1}{2}} + a_{j+\frac{1}{2}}^- \Delta t.$$

It is clear that under a sufficiently small time step, the non-smooth regions $I_{j+\frac{1}{2}}$ are separated due to the finite wave speeds. In this way, the width of Riemann fan is bounded by $(a_{j+\frac{1}{2}}^+ - a_{j+\frac{1}{2}}^-) \Delta t$, in contrast to the fixed width Δx in the fully-discrete counterpart.

Now we reconstruct a non-oscillatory, conservative, piecewise polynomial interpolant at $t = t^{n+1}$,

$$\tilde{w}^{n+1}(x) = \sum_j [\tilde{w}_j^n(x) \mathbf{1}_{[x_{j-\frac{1}{2}}^+, x_{j+\frac{1}{2}}^-]} + \tilde{w}_{j+\frac{1}{2}}^n(x) \mathbf{1}_{[x_{j+\frac{1}{2}}^-, x_{j+\frac{1}{2}}^+]}]. \quad (3.2.16)$$

The polynomials $\tilde{w}_j^{n+1}$ and $\tilde{w}_{j+\frac{1}{2}}^{n+1}$ conserve the cell averages, $\bar{w}_j^{n+1}$ and $\bar{w}_{j+\frac{1}{2}}^{n+1}$, over the smooth and non-smooth domains. The values of $\bar{w}_j^{n+1}$ and $\bar{w}_{j+\frac{1}{2}}^{n+1}$ are computed by integrating (3.2.1) over the rectangular domains, $[t^n, t^{n+1}] \times [x_{j-\frac{1}{2}}^+, x_{j+\frac{1}{2}}^-]$ and $[t^n, t^{n+1}] \times [x_{j+\frac{1}{2}}^-, x_{j+\frac{1}{2}}^+]$, respectively.

Finally, the cell average at t^{n+1} is computed with

$$\bar{u}_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{w}^{n+1}(x) dx.$$

Passing the limit $\Delta t \rightarrow 0$ yields a semi-discrete conservative scheme

$$\frac{d}{dt} \bar{u}_j(t) := \lim_{\Delta t \rightarrow 0} \frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} = - \frac{F_{j+\frac{1}{2}}(t) - F_{j-\frac{1}{2}}(t)}{\Delta x}, \quad (3.2.17a)$$

where the numerical fluxes, $F_{j+\frac{1}{2}}(t)$, are given by

$$F_{j+\frac{1}{2}}(t) := \frac{a_{j+\frac{1}{2}}^+ f(u_{j+\frac{1}{2}}^-) - a_{j+\frac{1}{2}}^- f(u_{j+\frac{1}{2}}^+)}{a_{j+\frac{1}{2}}^+ - a_{j+\frac{1}{2}}^-} + \frac{a_{j+\frac{1}{2}}^+ a_{j+\frac{1}{2}}^-}{a_{j+\frac{1}{2}}^+ - a_{j+\frac{1}{2}}^-} (u_{j+\frac{1}{2}}^+ - u_{j+\frac{1}{2}}^-). \quad (3.2.17b)$$

There are different options to estimate the wave speeds $a_{j+\frac{1}{2}}^\pm$. The *KT scheme* [87] proposed

to use the spectral radius of flux Jacobian,

$$a_{j+\frac{1}{2}}^+ = -a_{j+\frac{1}{2}}^- = a_{j+\frac{1}{2}} := \max\{\rho(\frac{\partial f}{\partial u}(u_{j+\frac{1}{2}}^-)), \rho(\frac{\partial f}{\partial u}(u_{j+\frac{1}{2}}^+))\},$$

where $\rho(A)$ represents the spectral radius of a matrix A . Then the numerical flux can be expressed in terms of the Rusanov flux,

$$F_{j+\frac{1}{2}}(t) = F^{Rus}(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) := \frac{f(u_{j+\frac{1}{2}}^-) + f(u_{j+\frac{1}{2}}^+)}{2} - \frac{a_{j+\frac{1}{2}}}{2}(u_{j+\frac{1}{2}}^+ - u_{j+\frac{1}{2}}^-). \quad (3.2.18)$$

For strictly hyperbolic problems, the flux Jacobian $\frac{\partial f}{\partial u}$ has N distinct eigenvalues $\lambda_1 < \dots < \lambda_N$. The *semi-discrete central-upwind scheme* [86] employs a more accurate estimation of the wave speeds,

$$\begin{aligned} a_{j+\frac{1}{2}}^+ &= \max\{\lambda_N(\frac{\partial f}{\partial u}(u_{j+\frac{1}{2}}^-)), \lambda_N(\frac{\partial f}{\partial u}(u_{j+\frac{1}{2}}^+)), 0\}, \\ a_{j+\frac{1}{2}}^- &= \min\{\lambda_1(\frac{\partial f}{\partial u}(u_{j+\frac{1}{2}}^-)), \lambda_1(\frac{\partial f}{\partial u}(u_{j+\frac{1}{2}}^+)), 0\}. \end{aligned} \quad (3.2.19)$$

In this way, the scheme reduces the numerical dissipation by employing the HLL flux [55].

In the computations, the set of ODEs (3.2.17) is integrated with an appropriate ODE solver. To preserve the overall high order of accuracy as well as the non-oscillatory property, one may apply a higher order *Strong Stability Preserving* (SSP) Runge-Kutta method, see e.g. [49, 50, 132].

The semi-discrete central schemes retain the advantage of being Riemann-solver-free, hence the component-wise extension is allowed when solving systems of equations. In multidimensional problems, the second order constructions can be extended in a dimension-by-dimension manner. Third order extensions have also been derived with the help of piecewise parabolic interpolants, we refer the details to [85, 86].

3.3 Revisiting the smooth van Albada slope limiter

The key to success of MUSCL schemes is a proper construction of slope limiter. Most of the existing slope limiters are variants of the minmod limiter (3.2.5). They share an important common feature: near the local extrema ($\Delta \bar{u}_{j-\frac{1}{2}} \cdot \Delta \bar{u}_{j+\frac{1}{2}} < 0$) the approximate slope must vanish, $u'_j = 0$, and thus the polynomial reconstruction degenerates to first order piecewise constant interpolation. We call such feature the *minmod structure*.

In the numerous studies of limiter schemes over the last decades, minmod-type limiters have been frequently taken for most discussions. However, there is a slope limiter that is not of minmod-type – the smooth van Albada limiter,

$$\psi_{VA}^\epsilon(a, b) = \frac{(a^2 + \epsilon^2)b + (b^2 + \epsilon^2)a}{a^2 + b^2 + 2\epsilon^2}, \quad \epsilon^2 = O(\Delta x^3). \quad (3.3.1)$$

This limiter was first proposed by van Albada, van Leer and Robert [147] in 1982 and has been shown to generate non-oscillatory solutions in many computations e.g. [102, 121, 124], but its theoretical stability has not been rigorously proved. In this section, we prove that the van Albada limiter generates second order, TVD solutions as applied to central schemes, despite the lack of minmod structure.

To get more insight, the limiter (3.3.1) can be written as [149]

$$\psi_{VA}^\epsilon(a, b) = \frac{a+b}{2} \left(1 - \frac{(a-b)^2}{a^2 + b^2 + 2\epsilon^2} \right).$$

In the smooth regions where $a \approx b$, the limiter tends to recover the second order central finite differencing, $\frac{a+b}{2}$. Across the discontinuities, however, the averaged slope is biased to the smallest value among the two one-sided slopes. These mechanisms are expected to ensure the second order accuracy and prevent undesirable numerical oscillations. We will provide rigorous proofs for these advantages. For convenience of discussion, we will consider the simplified

version without ϵ

$$\psi_{VA}(a, b) = \frac{a^2b + ab^2}{a^2 + b^2}, \quad a^2 + b^2 \neq 0. \quad (3.3.2)$$

Indeed, as reported in [147] the computed solutions are not sensitive to the specific value of ϵ .

3.3.1 The second order accuracy

We show that the van Albada limiter (3.3.2) ensures a second order of accuracy in space. We first consider the reconstruction based on the *point values*, $u_j := u(x_j)$. Denote the left- and the right- sided slopes by

$$\partial_x^- u_j := \frac{u_j - u_{j-1}}{\Delta x} \quad \text{and} \quad \partial_x^+ u_j := \frac{u_{j+1} - u_j}{\Delta x}.$$

The numerical derivative constructed with the van Albada limiter (3.3.2) is given by

$$\frac{u'_j}{\Delta x} = \frac{(\partial_x^- u_j)(\partial_x^+ u_j)^2 + (\partial_x^+ u_j)(\partial_x^- u_j)^2}{(\partial_x^+ u_j)^2 + (\partial_x^- u_j)^2}.$$

Taylor expansion yields

$$\frac{u'_j}{\Delta x} = \partial_x u_j + \left(\frac{1}{6} \partial_x^3 u_j - \frac{1}{2} \frac{(\partial_x^2 u_j)^2}{\partial_x u_j} \right) \Delta x^2 + O(\Delta x^3) \quad (3.3.3)$$

The second order condition (3.2.4) is satisfied at the non-critical points, $\partial_x u_j \neq 0$. In the framework of finite volume methods, the slope reconstruction is based on the *cell averages*, $\bar{u}_j := \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x) dx$. Then the Taylor expansion is slightly modified,

$$\frac{u'_j}{\Delta x} = \partial_x u_j + \left(\frac{1}{6} \partial_x^3 u_j + \frac{1}{24} \partial_x^4 u_j - \frac{1}{2} \frac{\partial_x^2 u_j}{\partial_x u_j} \right) \Delta x^2 + O(\Delta x^3) \quad (3.3.4)$$

For both point-value-based and cell-average-based reconstructions, the numerical derivative, $u'_j/\Delta x$, approximates the exact derivative, $\partial_x u_j$, with order $O(\Delta x^2)$ in the non-critical re-

gions. Near the critical points, the second order condition (3.2.4) is preserved up to $\partial_x u \approx \Delta x$. Hence it is expected that the van Albada limiter introduces less dissipation than the minmod limiter at discontinuities and smooth extrema. Their numerical performances will be compared in section 3.4.

3.3.2 The non-oscillatory property

We are going to show that the van Albada limiter, as applied to the central schemes, generates non-oscillatory solutions. We begin with the following lemma.

Lemma 3.3.1 *The approximate slope $u'_j = \psi_{VA}(\Delta \bar{u}_{j-\frac{1}{2}}, \Delta \bar{u}_{j+\frac{1}{2}})$ satisfies the following estimates*

$$\frac{1 - \sqrt{2}}{2} \leq \frac{u'_j}{\Delta \bar{u}_{j \pm \frac{1}{2}}} \leq \frac{1 + \sqrt{2}}{2}, \quad (3.3.5a)$$

$$\frac{|u'_{j+1} - u'_j|}{|\Delta \bar{u}_{j+\frac{1}{2}}|} \leq \sqrt{2}. \quad (3.3.5b)$$

Proof. Denote $r := \frac{\Delta \bar{u}_{j-\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}}$, then we can write

$$u'_j = \frac{r^2 + r}{r^2 + 1} \Delta \bar{u}_{j+\frac{1}{2}} = \frac{r + 1}{r^2 + 1} \Delta \bar{u}_{j-\frac{1}{2}}.$$

We will show that $\frac{1 - \sqrt{2}}{2} \leq \frac{r^2 + r}{r^2 + 1}, \frac{r + 1}{r^2 + 1} \leq \frac{1 + \sqrt{2}}{2}$.

Assume that $\frac{r^2 + r}{r^2 + 1}$ is bounded from below and above by m and $M \in \mathbb{R}$ respectively,

$$m \leq \frac{r^2 + r}{r^2 + 1} \leq M, \quad \forall r \in \mathbb{R}.$$

This can be equivalently expressed as

$$\begin{cases} (M - 1)r^2 - r + M \geq 0 \\ (m - 1)r^2 - r + m \leq 0 \end{cases}, \quad \forall r \in \mathbb{R}. \quad (3.3.6)$$

The upper bound, M , should satisfy the conditions

$$\begin{cases} M > 1 \\ \Delta = 1 - 4M(M - 1) \leq 0 \end{cases}.$$

Solving these inequalities yields $M \geq \frac{1 + \sqrt{2}}{2}$ and hence $\frac{1 + \sqrt{2}}{2}$ gives the supremum of $\frac{r^2 + r}{r^2 + 1}$.

On the other hand, the lower bound, m , should satisfy

$$\begin{cases} m < 1 \\ \Delta = 1 - 4m(m - 1) \leq 0 \end{cases}.$$

These conditions yield $m \leq \frac{1 - \sqrt{2}}{2}$ and hence $\frac{1 - \sqrt{2}}{2}$ gives the infimum of $\frac{r^2 + r}{r^2 + 1}$. Combining these results, we have

$$\frac{1 - \sqrt{2}}{2} \leq \frac{u'_j}{\Delta \bar{u}_{j+\frac{1}{2}}} = \frac{r^2 + r}{r^2 + 1} \leq \frac{1 + \sqrt{2}}{2}.$$

With similar argument, we can show that $\frac{1 - \sqrt{2}}{2} \leq \frac{r + 1}{r^2 + 1} \leq \frac{1 + \sqrt{2}}{2}$ and therefore

$$\frac{1 - \sqrt{2}}{2} \leq \frac{u'_j}{\Delta \bar{u}_{j-\frac{1}{2}}} = \frac{r + 1}{r^2 + 1} \leq \frac{1 + \sqrt{2}}{2}.$$

We complete the proof of (3.3.5a). The second estimate (3.3.5b) is a direct consequence of (3.3.5a). $\square$

With the help of Lemma 3.3.1, we discuss the non-oscillatory property of the fully-discrete and the semi-discrete schemes separately.

3.3.2.1 The fully-discrete scheme

We consider the fully-discrete NT scheme (3.2.10). The scheme can be written in a conservative form [107]

$$\bar{u}_{j+\frac{1}{2}}^{n+1} = \frac{\bar{u}_j^n + \bar{u}_{j+1}^n}{2} - \lambda(g_{j+1} - g_j), \quad (3.3.7)$$

where the modified numerical flux, g_j , is given by

$$g_j = f(u_j^{n+\frac{1}{2}}) + \frac{1}{8\lambda} u_j'. \quad (3.3.8)$$

The NT scheme combined with the van Albada limiter (3.3.2) can be shown to be TVD following similar approach in [107]. The proof is based on the following lemma.

Lemma 3.3.2 ([107]) *The scheme (3.2.10) is TVD if the numerical flux, g_j , satisfies the following generalized CFL condition*

$$\lambda \left| \frac{\Delta g_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}^n} \right| \leq \frac{1}{2}, \quad \Delta g_{j+\frac{1}{2}} := g_{j+1} - g_j. \quad (3.3.9)$$

Proof. The proof follows the lines of [52]. By (3.3.7), the difference $\bar{u}_{j+\frac{1}{2}}^{n+1} - \bar{u}_{j-\frac{1}{2}}^{n+1}$ is equal to

$$\bar{u}_{j+\frac{1}{2}}^{n+1} - \bar{u}_{j-\frac{1}{2}}^{n+1} = \left(\frac{1}{2} - \lambda \frac{\Delta g_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}^n} \right) \Delta \bar{u}_{j+\frac{1}{2}} + \left(\frac{1}{2} + \lambda \frac{\Delta g_{j-\frac{1}{2}}}{\Delta \bar{u}_{j-\frac{1}{2}}^n} \right) \Delta \bar{u}_{j-\frac{1}{2}}.$$

Condition (3.3.10) implies that the coefficients in the parenthesis are positive. Hence,

$$\begin{aligned} \sum_j |\bar{u}_{j+\frac{1}{2}}^{n+1} - \bar{u}_{j-\frac{1}{2}}^{n+1}| &\leq \sum_j \left(\frac{1}{2} - \lambda \frac{\Delta g_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}^n} \right) |\Delta \bar{u}_{j+\frac{1}{2}}| + \left(\frac{1}{2} + \lambda \frac{\Delta g_{j-\frac{1}{2}}}{\Delta \bar{u}_{j-\frac{1}{2}}^n} \right) |\Delta \bar{u}_{j-\frac{1}{2}}| \\ &= \sum_j \left(\frac{1}{2} - \lambda \frac{\Delta g_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}^n} \right) |\Delta \bar{u}_{j+\frac{1}{2}}| + \left(\frac{1}{2} + \lambda \frac{\Delta g_{j+\frac{1}{2}}}{\Delta \bar{u}_{j-\frac{1}{2}}^n} \right) |\Delta \bar{u}_{j+\frac{1}{2}}| \\ &= \sum_j |\Delta \bar{u}_{j+\frac{1}{2}}|. \end{aligned}$$

This concludes

$$TV(\bar{u}^{n+1}) \leq TV(\bar{u}^n).$$

□

Equipped with Lemma 3.3.1 and 3.3.2, we prove the NT scheme is TVD.

Theorem 3.3.3 *Let the numerical slope u'_j be constructed with the van Albada limiter (3.3.2), and let the flux numerical derivative be chosen by $f'_j = a(\bar{u}_j^n)u'_j$ with $a(u) := f'(u)$. Assume that the following CFL condition is satisfied*

$$\lambda \max_j |a(\bar{u}_j)| \leq \frac{-1 + \sqrt{\frac{3}{2} + \frac{3\sqrt{2}}{4}}}{1 + \sqrt{2}} \approx 0.24. \quad (3.3.10)$$

Then the NT scheme (3.3.7) is TVD.

Proof. Denote $cfl := \lambda \max_u |a(u)|$. By (3.3.8) we have

$$\lambda \left| \frac{\Delta g_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}} \right| \leq \underbrace{\lambda \left| \frac{f(u_{j+1}^{n+\frac{1}{2}}) - f(u_j^{n+\frac{1}{2}})}{u_{j+1}^{n+\frac{1}{2}} - u_j^{n+\frac{1}{2}}} \right|}_I \underbrace{\left| \frac{u_{j+1}^{n+\frac{1}{2}} - u_j^{n+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}} \right|}_{II} + \underbrace{\frac{1}{8} \left| \frac{\Delta u'_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}} \right|}_{III}. \quad (3.3.11)$$

By using the definition, the first term on the right of (3.3.11) is bounded with

$$\lambda \left| \frac{f(u_{j+1}^{n+\frac{1}{2}}) - f(u_j^{n+\frac{1}{2}})}{u_{j+1}^{n+\frac{1}{2}} - u_j^{n+\frac{1}{2}}} \right| \leq \lambda \max_u |a(u)| := cfl. \quad (3.3.12)$$

Notice that (3.3.5a) implies $\left| \frac{u'_j}{\Delta \bar{u}_{j \pm \frac{1}{2}}} \right| \leq \frac{1 + \sqrt{2}}{2}$, we can estimate the second term on the right of (3.3.11),

$$\begin{aligned} \left| \frac{u_{j+1}^{n+\frac{1}{2}} - u_j^{n+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}} \right| &\leq 1 + \frac{\lambda}{2} \left| \frac{\Delta f'_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}} \right| \\ &\leq 1 + \frac{\lambda}{2} \frac{|a_{j+1} u'_{j+1}| + |a_j u'_j|}{|\Delta \bar{u}_{j+\frac{1}{2}}|} \\ &\leq 1 + \frac{\lambda \max_u |a(u)|}{2} \frac{|u'_{j+1}| + |u'_j|}{|\Delta \bar{u}_{j+\frac{1}{2}}|} \\ &\leq 1 + \frac{1 + \sqrt{2}}{2} cfl. \end{aligned} \quad (3.3.13)$$

As a direct consequence of (3.3.5b), the third term on the right of (3.3.11) does not exceed

$$\frac{1}{8} \left| \frac{\Delta u'_{j+\frac{1}{2}}}{\Delta \bar{u}_{j+\frac{1}{2}}} \right| \leq \frac{\sqrt{2}}{8}. \quad (3.3.14)$$

Using (3.3.12), (3.3.13) and (3.3.14), the TVD condition (3.3.9) boils down to

$$cfl \left(1 + \frac{1 + \sqrt{2}}{2} cfl \right) + \frac{\sqrt{2}}{8} \leq \frac{1}{2}.$$

Solving this quadratic inequality yields

$$0 \leq cfl \leq \frac{-1 + \sqrt{\frac{3}{2} + \frac{3\sqrt{2}}{4}}}{1 + \sqrt{2}}. \quad (3.3.15)$$

The condition (3.3.10) is recovered. □

Remark 3.3.4 *It is noticed that the CFL condition (3.3.10) is not sharp. Nevertheless, we point out that (3.3.10) merely serves as a theoretically sufficient condition. As shown by numerical experiments in Section 3.4, the computed solutions remain non-oscillatory as long as $cfl \leq \frac{1}{2}$.*

3.3.2.2 The semi-discrete scheme

The TVD stability of the van Albada limiter as applied to the non-staggered semi-discrete schemes (3.2.17) can be shown following the lines of [112, 140]. The key ingredient is to represent the scheme in an appropriate incremental form which meets a certain positive condition. We prove the following result with the help of Lemma 3.3.1.

Theorem 3.3.5 *Let the states $u_{j+\frac{1}{2}}^{\pm}$ be computed from the MUSCL reconstruction (3.2.15), with the numerical slope u'_j constructed with the van Albada limiter (3.3.2). Consider a generalized MUSCL scheme*

$$\frac{d}{dt} \bar{u}_j(t) = -\frac{1}{\Delta x} [F(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) - F(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+)], \quad (3.3.16)$$

and let the numerical flux $F(\cdot, \cdot)$ be monotone and consistent,

$$\frac{\partial}{\partial u} F(u, v) \geq 0 \quad \text{and} \quad \frac{\partial}{\partial v} F(u, v) \leq 0, \quad (3.3.17a)$$

$$F(u, u) = f(u). \quad (3.3.17b)$$

Then the scheme (3.3.16) is TVD,

$$\frac{d}{dt} [TV(\bar{u}(t))] \leq 0.$$

Proof. We write the scheme (3.3.16) in the incremental form

$$\frac{d}{dt} \bar{u}_j = -\frac{1}{\Delta x_j} C_{j-\frac{1}{2}} \Delta \bar{u}_{j-\frac{1}{2}} + \frac{1}{\Delta x_j} D_{j+\frac{1}{2}} \Delta \bar{u}_{j+\frac{1}{2}}, \quad (3.3.18)$$

where

$$\begin{aligned} C_{j-\frac{1}{2}} &= \frac{1}{\Delta \bar{u}_{j-\frac{1}{2}}} [F(u_{j+\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+) - F(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+)], \\ D_{j+\frac{1}{2}} &= -\frac{1}{\Delta \bar{u}_{j+\frac{1}{2}}} [F(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) - F(u_{j+\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+)]. \end{aligned} \quad (3.3.19)$$

We denote

$$\begin{aligned} s_{j+\frac{1}{2}} &:= \text{sign}(\Delta \bar{u}_{j+\frac{1}{2}}) = \begin{cases} 1 & \Delta \bar{u}_{j+\frac{1}{2}} > 0 \\ \pm 1 & \Delta \bar{u}_{j+\frac{1}{2}} = 0 \\ -1 & \Delta \bar{u}_{j+\frac{1}{2}} < 0 \end{cases}, \\ \chi_j &:= 1 - s_{j-\frac{1}{2}} s_{j+\frac{1}{2}} = \begin{cases} 2, & \Delta \bar{u}_{j-\frac{1}{2}} \cdot \Delta \bar{u}_{j+\frac{1}{2}} \leq 0 \\ 0, & \Delta \bar{u}_{j-\frac{1}{2}} \cdot \Delta \bar{u}_{j+\frac{1}{2}} > 0 \end{cases}. \end{aligned}$$

Forward differencing of (3.3.18) gives

$$\begin{aligned} \frac{d}{dt} \Delta \bar{u}_{j+\frac{1}{2}} &= \left(\frac{1}{\Delta x_{j+1}} D_{j+\frac{3}{2}} \Delta \bar{u}_{j+\frac{3}{2}} - \frac{1}{\Delta x_{j+1}} C_{j+\frac{1}{2}} \Delta \bar{u}_{j+\frac{1}{2}} \right) \\ &\quad - \left(\frac{1}{\Delta x_j} D_{j+\frac{1}{2}} \Delta \bar{u}_{j+\frac{1}{2}} - \frac{1}{\Delta x_j} C_{j-\frac{1}{2}} \Delta \bar{u}_{j-\frac{1}{2}} \right). \end{aligned} \quad (3.3.20)$$

Multiply (3.3.20) by $s_{j+\frac{1}{2}}$ and sum by parts, we have

$$\begin{aligned}
\frac{d}{dt}TV(\bar{u}) &= \sum_j s_{j+\frac{1}{2}} \frac{d}{dt} \Delta \bar{u}_{j+\frac{1}{2}} \\
&= - \sum_j \frac{1}{\Delta x} [(s_{j-\frac{1}{2}} - s_{j+\frac{1}{2}}) C_{j-\frac{1}{2}} \Delta \bar{u}_{j-\frac{1}{2}} + (s_{j+\frac{1}{2}} - s_{j-\frac{1}{2}}) D_{j+\frac{1}{2}} \Delta \bar{u}_{j+\frac{1}{2}}] \\
&= - \sum_j \frac{\chi_j}{\Delta x} [C_{j-\frac{1}{2}} |\Delta \bar{u}_{j-\frac{1}{2}}| + D_{j+\frac{1}{2}} |\Delta \bar{u}_{j+\frac{1}{2}}|].
\end{aligned}$$

Here we use the property that $s_{j+\frac{1}{2}}^2 \equiv 1$. Since $\chi_j \geq 0$, it suffices to show that $C_{j-\frac{1}{2}}, D_{j+\frac{1}{2}} \geq 0$.

Indeed, we find that

$$\begin{aligned}
C_{j-\frac{1}{2}} &= \partial_u F(\tilde{u}_1, u_{j-\frac{1}{2}}^+) (1 + \frac{u'_j - u'_{j-1}}{2\Delta \bar{u}_{j-\frac{1}{2}}}), \\
D_{j+\frac{1}{2}} &= -\partial_v F(u_{j+\frac{1}{2}}^-, \tilde{u}_2) (1 - \frac{u'_{j+1} - u'_j}{2\Delta \bar{u}_{j+\frac{1}{2}}}),
\end{aligned} \tag{3.3.21}$$

where $\tilde{u}_1$ is between $u_{j-\frac{1}{2}}^-$ and $u_{j+\frac{1}{2}}^-$, $\tilde{u}_2$ is between $u_{j-\frac{1}{2}}^+$ and $u_{j+\frac{1}{2}}^+$. From (3.3.5b) we know that the terms in the parentheses are positive. Hence the non-negativity of $C_{j-\frac{1}{2}}$ and $D_{j+\frac{1}{2}}$ follows from (3.3.17a). $\square$

The semi-discrete central scheme (3.2.17) can be considered a generalized MUSCL scheme, the associated numerical flux (3.2.17b) can be expressed in the form $F_{j+\frac{1}{2}}(t) := F(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+)$, with

$$F(u, v) = \frac{a^+ f(u) - a^- f(v)}{a^+ - a^-} + \frac{a^+ a^-}{a^+ - a^-} (v - u).$$

The monotonicity of $F(\cdot, \cdot)$ is clear since $a^- \leq f'(u), f'(v) \leq a^+$. Hence the TVD stability of the van Albada limiter as applied to the semi-discrete central schemes (3.2.17) follows from Theorem 3.3.5.

Remark 3.3.6 *The semi-discrete scheme (3.2.17) (or equivalently (3.3.16)) advanced with forward Euler time-stepping reads*

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \lambda (F(u_{j+\frac{1}{2}}^-, u_{j+\frac{1}{2}}^+) - F(u_{j-\frac{1}{2}}^-, u_{j-\frac{1}{2}}^+)), \tag{3.3.22}$$

where $u_{j+\frac{1}{2}}^+ := \bar{u}_j^n - \frac{1}{2}u'_{j+1}$ and $u_{j+\frac{1}{2}}^- := \bar{u}_j^n + \frac{1}{2}u'_j$. With similar incremental representation in Theorem 3.3.5, it can be proved that the van Albada limiter as applied to the fully-discrete non-staggered scheme (3.3.22) generates TVD solutions. Applying a higher order SSP time integration method would achieve a higher order of accuracy in time without deteriorating the stability.

3.4 Numerical examples

We examine the performance of MUSCL schemes in several test cases. The discussions of this section include two aspects. The first aspect is to verify the advantage of the van Albada limiter (3.3.1) over the classical minmod limiter (3.2.5). Examples for conservation laws and incompressible Euler equations will be presented. We take the small bias $\epsilon = \Delta x^3$ for van Albada limiter in all the computations. The second aspect is to validate the application of MUSCL schemes in collective dynamics simulations. In particular, we will illustrate the hydrodynamic alignment of single-species in one- and two-dimensions through numerical experiments.

3.4.1 Conservation laws

We verify the high resolution and the non-oscillatory property of limiter (3.3.1) as applied to the fully discrete NT scheme (3.2.10) and the semi-discrete central upwind scheme (see (3.2.17) and (3.2.19)). Several test cases for linear and nonlinear conservation laws are presented.

In all the simulations of the NT scheme, the numerical slope of flux, f'_j is evaluated with the exact flux Jacobian (3.2.12), and the CFL number is taken to be 0.45. The implementation of the semi-discrete scheme requires a high order time discretization. We use the third order, explicit SSPRK3 method for the time integration of method-of-lines ODEs (3.2.17a) with the CFL number chosen to be 0.7.

3.4.1.1 One-dimensional linear advection equation

We consider the one-dimensional linear advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

with periodic boundary conditions over the domain $[0, 2\pi]$. The initial data includes a $\sin^4(\pi x)$ distribution, a step function and a function. We compute the solution after one-period of revolution under $N = 400$ equidistant mesh grids. Figure 3.4.1 displays the solutions computed with the NT scheme. It is seen that both the van Albada limiter and the minmod limiter generate non-oscillatory solutions with comparable resolution at the shocks and smooth extrema. Figure 3.4.2 compares the solutions computed with different limiters as applied to the semi-discrete central-upwind scheme. In this case, applying the van Albada limiter greatly improves the sharpness of the solution at the peaks of the sinusoidal wave and the triangular wave.

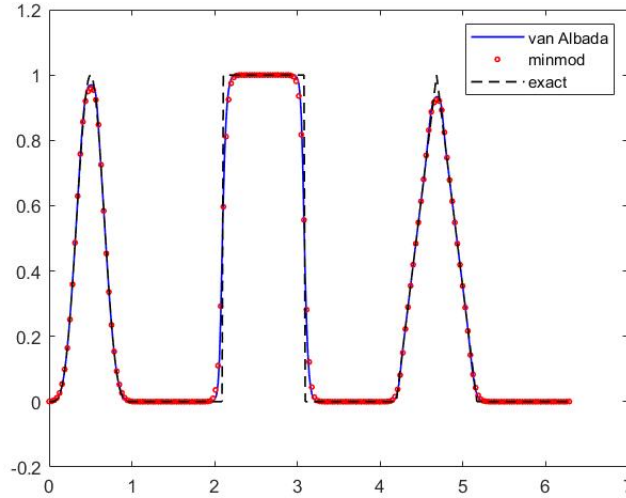


Figure 3.4.1: Linear advection, NT scheme, $N = 200$, periodic bc, $t = 2\pi$.

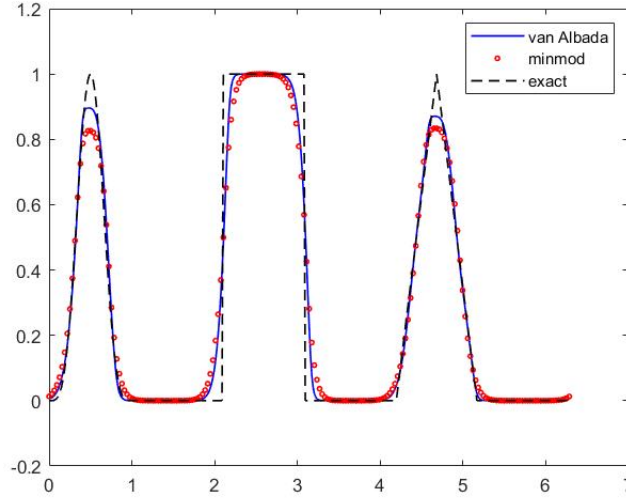


Figure 3.4.2: Linear advection, semi-discrete central-upwind scheme, $N = 200$, periodic bc, $t = 2\pi$.

3.4.1.2 One-dimensional Euler equations

We consider the one-dimensional Euler equations of gas dynamics

$$\frac{\partial}{\partial t} \begin{bmatrix} \rho \\ \rho u \\ E \end{bmatrix} + \frac{\partial}{\partial x} \begin{bmatrix} \rho u \\ \rho u^2 + p \\ (E + p)u \end{bmatrix} = 0,$$

where ρ , u , p are the density, velocity and pressure, E is the total energy per unit volume given by

$$E = \frac{p}{\gamma - 1} + \frac{\rho u^2}{2}, \quad \gamma = 1.4.$$

At first, we consider the Sod's shock tube problem. The computational domain is $[0, 1]$ with an interface at $x = 0.5$. The initial states on the two sides of the interface are given by

$$\begin{cases} p_L = 1.0, & p_R = 0.1 \\ \rho_L = 1.0, & \rho_R = 0.125 \\ u_L = 0.0, & u_R = 0.0 \end{cases}.$$

We compute the numerical solution with $N = 400$ mesh grids. Figures 3.4.3–3.4.4 show the solutions of density at $t = 0.2$ by different schemes and limiters. The results have shown the high resolution and the non-oscillatory property of the van Albada limiter.

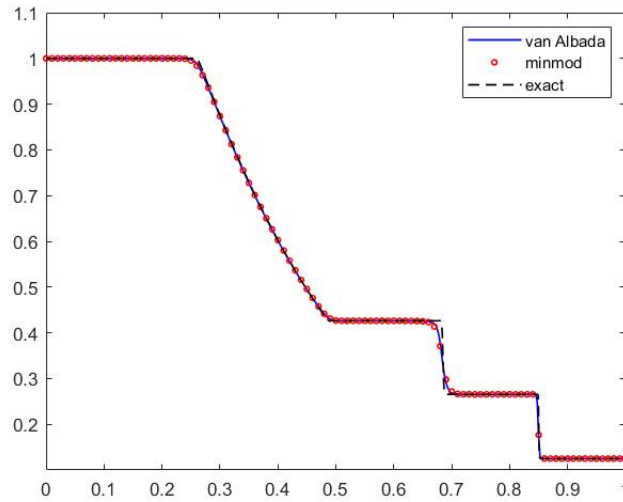


Figure 3.4.3: Sod's problem—density, NT scheme, $N = 400$, $t = 0.2$.

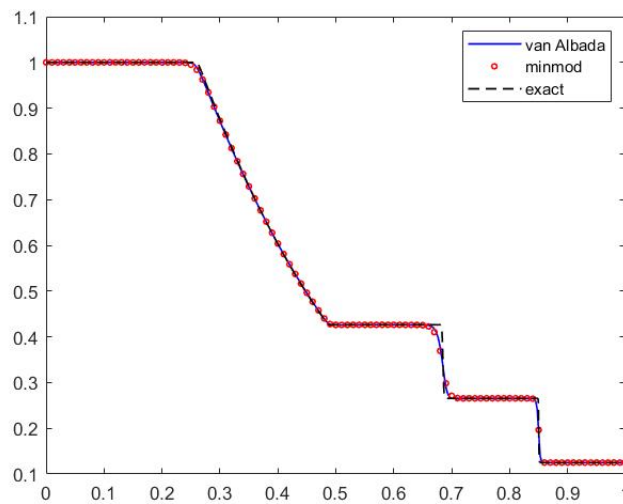


Figure 3.4.4: Sod's problem—density, semi-discrete central-upwind scheme, $N = 400$, $t = 0.2$.

The second test case is the Osher-Shu problem. The computational domain is $[-5, 5]$ with the interface placed at $x = -4$. The initial data is given by

$$\begin{cases} p_L = 10.33333, & p_R = 1.0 \\ \rho_L = 3.857143, & \rho_R = 1 + 0.2 \sin(5x) \\ u_L = 2.6293690, & u_R = 0.0 \end{cases} .$$

In this problem, the solution of density consists of a discontinuity and a smooth harmonic wave. Figure 3.4.3 and 3.4.6 present the solutions of density given by the NT scheme and the semi-discrete central-upwind scheme computed with $N = 600$ mesh grids at $t = 1.8$. It is seen that the solutions computed with the van Albada resolve the smooth extremum better than those computed with the minmod limiter, and the solutions do not introduce numerical oscillations at the shock even if the reconstructed slope does not have the minmod property.

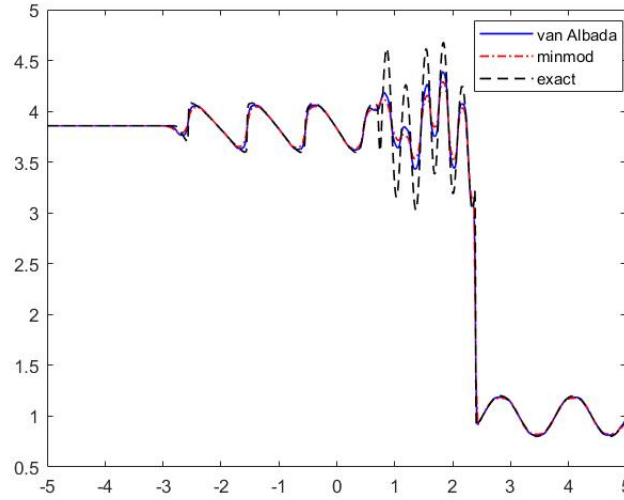


Figure 3.4.5: Osher-Shu problem–density, NT scheme, $N = 600$, $t = 1.8$.

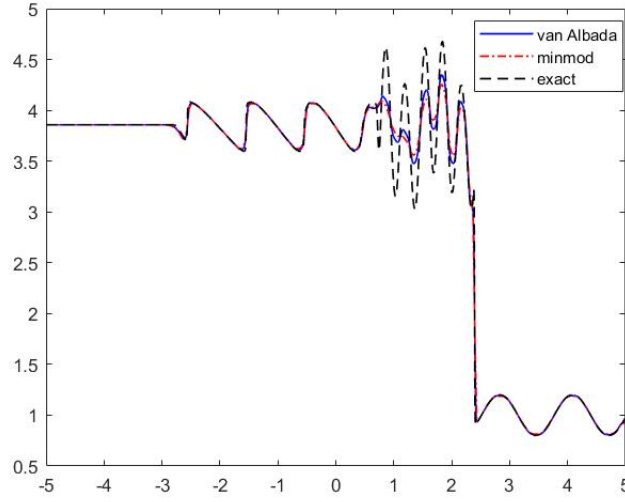


Figure 3.4.6: Osher-Shu problem—density, semidiscrete central-upwind scheme, $N = 600$, $t = 1.8$.

3.4.1.3 Two-dimensional Euler equations

We examine the performance of the van Albada limiter in multi-dimensions. We consider the double Mach reflection problem [156] for the two-dimensional Euler equations,

$$\frac{\partial}{\partial t} \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix} + \frac{\partial}{\partial x} \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (E + p)u \end{bmatrix} + \frac{\partial}{\partial y} \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (E + p)v \end{bmatrix} = 0, \quad p = (\gamma - 1) \cdot [E - \frac{\rho}{2}(u^2 + v^2)].$$

The problem is initiated with a Mach 10 oblique shock positioned at $(1/6, 0)$ which makes a 60° angel with x -axis. The computational domain is set to $[0, 4] \times [0, 1]$. The bottom boundary ($y = 0$) consists of a reflecting wall beginning at $x = 1/6$, the short region from $x = 0$ to $x = 1/6$ is imposed with the initial post-shock conditions. The left boundary ($x = 0$) is also assigned with the initial post-shock values, and at the right boundary ($x = 4$) all the gradients are set to zero. The values along the top boundary ($y = 1$) are set to describe the exact motion of the Mach 10 shock. We refer to [156] for more detailed descriptions. We compute the solution of density at

$t = 0.2$ with 480×120 mesh grids. Figure 3.4.7 shows the reference solution given by the 5th order WENO finite difference method [123].

Figures 3.4.8, 3.4.9 display the solutions computed with the NT scheme. It is clear from the results that the solution generated by the van Albada limiter has obviously better quality. The shock waves and the slip lines underneath are characterized much more sharply. The computation results of the semi-discrete scheme are presented in Figures 3.4.10, 3.4.11. The numerical scheme captures more information of the near-wall flow when used in conjunction with the van Albada limiter. These results confirm the advantage of the limiter (3.3.1) in terms of resolving multidimensional complex flow structures.

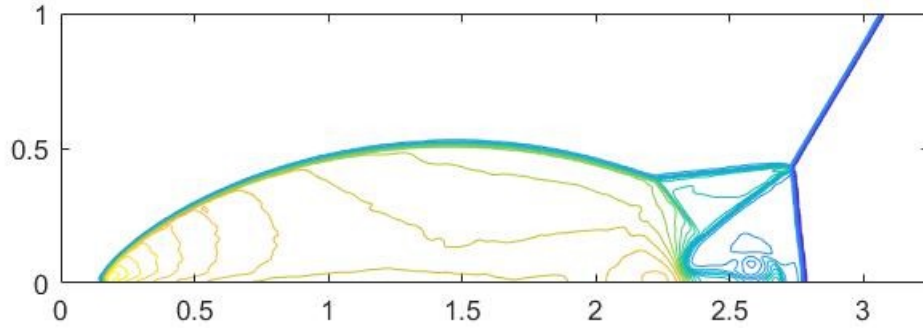


Figure 3.4.7: Double Mach reflection-density, 5th order WENO, $N_x = 480$, $N_y = 120$, $t = 0.2$.

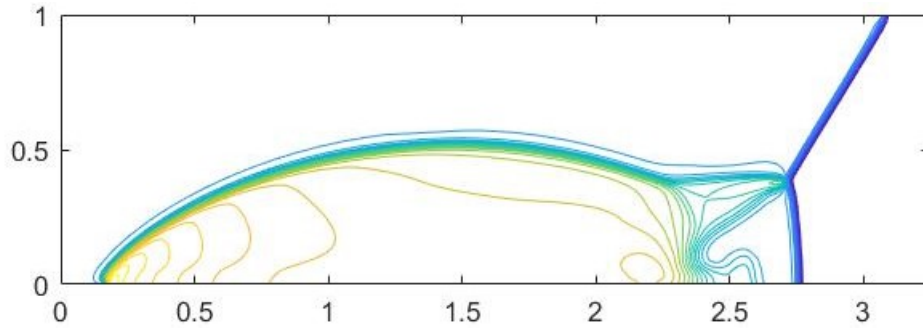


Figure 3.4.8: Double Mach reflection-density, NT scheme, minmod limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$.

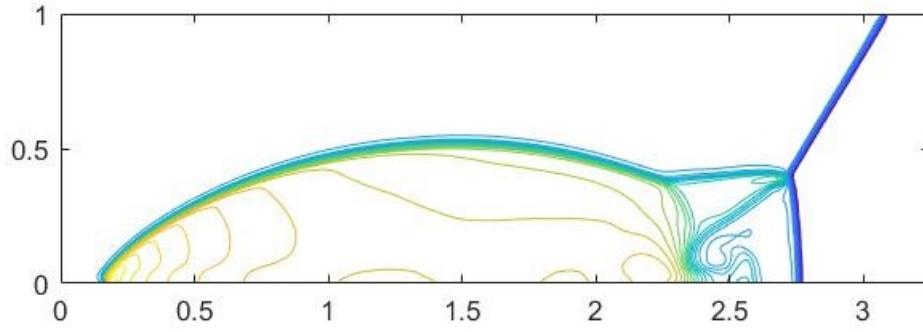


Figure 3.4.9: Double Mach reflection-density, NT scheme, van Albada limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$.

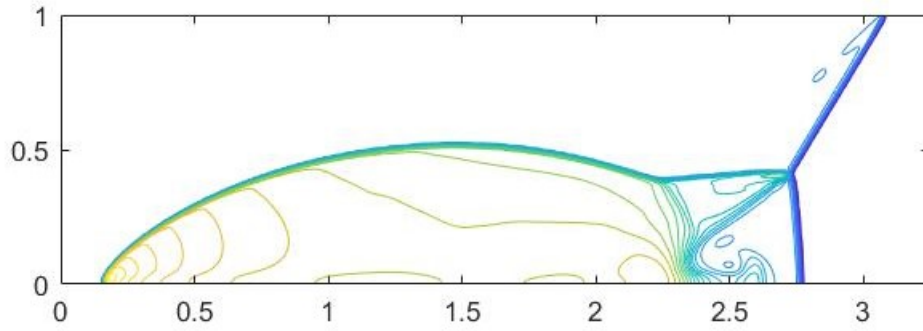


Figure 3.4.10: Double Mach reflection-density, semidiscrete central-upwind scheme, minmod limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$.

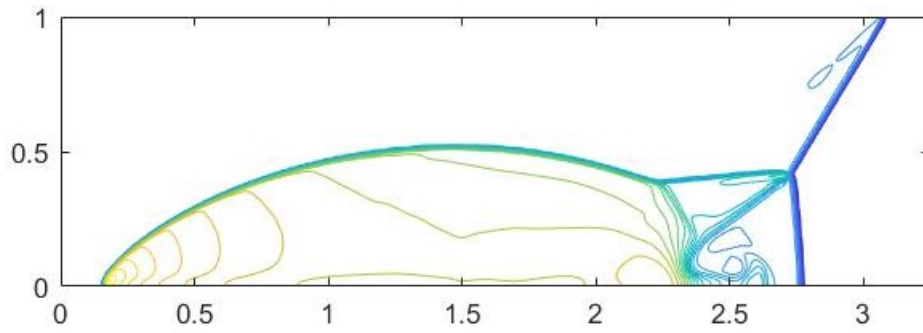


Figure 3.4.11: Double Mach reflection-density, semidiscrete central-upwind scheme, van Albada limiter, $N_x = 480$, $N_y = 120$, $t = 0.2$.

3.4.2 Incompressible Euler equations

We consider the two-dimensional double shear layer problem for incompressible Euler equations,

$$\begin{cases} \partial_t \mathbf{u} + \nabla \cdot (\mathbf{u} \otimes \mathbf{u}) = -\nabla p \\ \nabla \cdot \mathbf{u} = 0 \end{cases},$$

where p denotes the pressure and $\mathbf{u} = (u, v)$ is the velocity field. The initial data is given by

$$u_0(x, y) = \begin{cases} \tanh((y - \frac{\pi}{2})/\rho) & y \leq \pi \\ \tanh((\frac{3\pi}{2} - y)/\rho) & y > \pi \end{cases}, \quad v_0(x, y) = \delta \sin(x)$$

over the rectangular domain $[0, 2\pi] \times [0, 2\pi]$. We take $\rho = \frac{\pi}{15}$ and $\delta = 0.05$. The equations are solved with the finite difference projection method (see e.g. [33, 158]), which consists of a predictor step and a corrector step

$$\begin{aligned} \frac{\mathbf{u}^* - \mathbf{u}^n}{\Delta t} + \nabla \cdot (\mathbf{u}^n \otimes \mathbf{u}^n) &= 0, \\ \frac{\mathbf{u}^{n+1} - \mathbf{u}^*}{\Delta t} + \nabla p^{n+1} &= 0 \quad \text{with} \quad \nabla \cdot \mathbf{u}^{n+1} = 0. \end{aligned}$$

The divergence free constraint, $\nabla \cdot \mathbf{u}^{n+1} = 0$, leads to a Poisson equation for pressure

$$-\Delta p^{n+1} = -\frac{1}{\Delta t} \nabla \cdot \mathbf{u}^*.$$

In the predictor step, we reconstruct the convection fluxes, $\mathbf{u}^n \otimes \mathbf{u}^n$, with the Rusanov-type flux (3.2.18) coupled with the second order MUSCL reconstructions (3.2.15). In the corrector step, assuming the uniform mesh size is employed, $\Delta x = \Delta y = h$, the gradient, divergence and Laplace operators are approximated with the second order central difference operators,

$$\nabla^h \phi_{i,j} = \begin{pmatrix} \frac{\phi_{i+1,j} - \phi_{i-1,j}}{2\Delta x} \\ \frac{\phi_{i,j+1} - \phi_{i,j-1}}{2\Delta y} \end{pmatrix},$$

$$\begin{aligned}\nabla^h \cdot \mathbf{u}_{i,j} &= \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} + \frac{v_{i+1,j} - v_{i-1,j}}{2\Delta y}, \quad \mathbf{u} = (u, v), \\ \Delta^h \phi_{i,j} &= \frac{\phi_{i+2,j} - 2\phi_{i,j} + \phi_{i-2,j}}{4\Delta x^2} + \frac{\phi_{i,j+2} - 2\phi_{i,j} + \phi_{i,j-2}}{4\Delta y^2}.\end{aligned}$$

To achieve a higher order of accuracy in time, we apply the explicit SSPRK3 method for time marching. The time step is taken according to the CFL condition

$$\Delta t \max_{i,j} \left(\frac{|u_{i,j}|}{\Delta x} + \frac{|v_{i,j}|}{\Delta y} \right) \leq 0.5.$$

Figure 3.4.12 presents the solutions of vorticity at $t = 8$. We observe that the solution contours given by the van Albada limiter are more concentrated than those by the minmod limiter. Meanwhile, Figure 3.4.13 compares the solutions of v along the vertical line $x = \pi$. It is observed that the van Albada limiter has clear advantage over the minmod limiter in terms of capturing high-frequency extrema.

3.4.3 Hydrodynamic alignment of single-species

Now we apply the MUSCL scheme, combined with the van Albada limiter (3.3.1), to simulate the long time behavior of the pressure-less single-species hydrodynamics

$$\frac{\partial}{\partial t} \begin{bmatrix} \rho \\ \rho \mathbf{u} \end{bmatrix} + \nabla \cdot \begin{bmatrix} \rho \mathbf{u} \\ \rho \mathbf{u} \otimes \mathbf{u} \end{bmatrix} = \begin{bmatrix} 0 \\ \int \phi(\mathbf{x}, \mathbf{x}') (\mathbf{u}' - \mathbf{u}) \rho \rho' d\mathbf{x}' \end{bmatrix}.$$

The CFL condition for the time step Δt takes the form

$$\Delta t \max_{i,j} \left(\frac{|u_{i,j}|}{\Delta x} + \frac{|v_{i,j}|}{\Delta y} \right) \leq CFL.$$

In the following simulations, the convection term is reconstructed with the Rusanov-type numerical flux. Time marching is implemented with the explicit SSPRK2 method under $CFL = 0.7$.

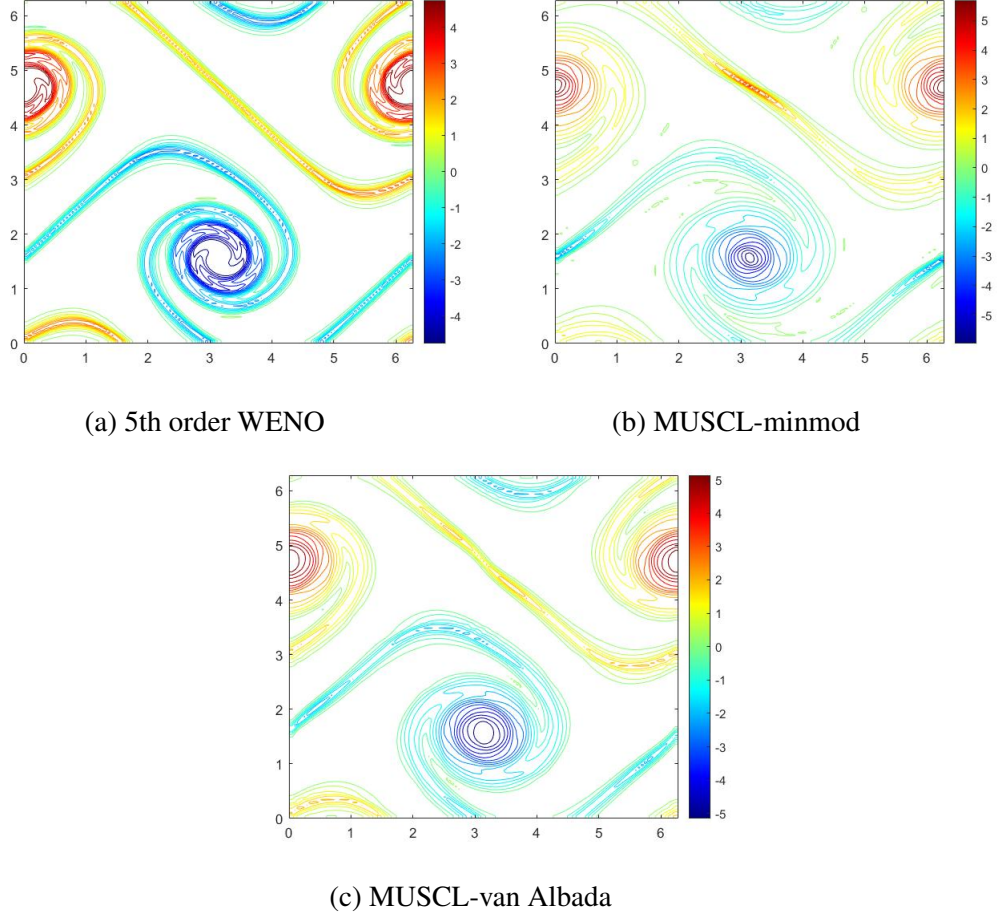


Figure 3.4.12: Double shear layer, contours of vorticity. $N_x = N_y = 128$, $CFL = 0.5$, $t = 8$.

3.4.3.1 Long-range communication

We study the hydrodynamics driven by the fat-tail kernel

$$\phi(\mathbf{x}, \mathbf{x}') = \frac{1}{|\mathbf{x} - \mathbf{x}'|^\theta}, \quad \theta > 0.$$

At first, we consider the one-dimensional dynamics subject to compactly supported initial data,

$$\rho_0(x) = 1, \quad u_0(x) = 1 + x, \quad -0.5 < x < 0.5. \quad (3.4.1)$$

The computation domain is taken to $[-5, 5]$ with periodic boundaries. Figure 3.4.14 displays the

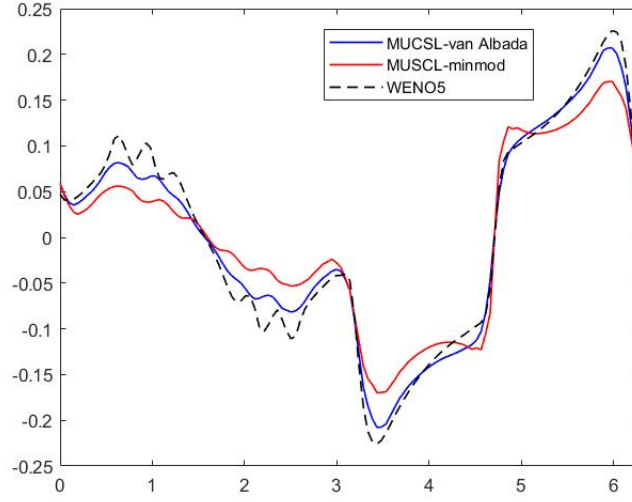


Figure 3.4.13: Double shear layer, solutions of v along $x = \pi$, $N_x = N_y = 128$, $CFL = 0.5$, $t = 8$.

histories of energy fluctuation,

$$\delta\mathcal{E}(t) = \int |\mathbf{u}(t, \mathbf{x}) - \mathbf{u}_\infty|^2 \rho(t, \mathbf{x}) d\mathbf{x},$$

up to $T = 5$ under different values of θ . The results recover the exponential decay for $\theta < 1$ and the polynomial decay for $\theta = 1$, as expected by Corollary 2.1.6. Interestingly, in our test case the fluctuation seems to decay with a polynomial rate for $\theta = 1.5$. In general, the emergence of alignment for $\theta > 1$ still lacks proof.

We also consider a two-dimensional test case with initial configuration

$$\rho_0(x, y) = 1, \quad u_0(x, y) = 1 + x, \quad v(x, y) = 1 + y, \quad -0.5 < x < 0.5, \quad -0.5 < y < 0.5. \quad (3.4.2)$$

The variation of energy fluctuation is presented in Figure 3.4.15. Again, the results verify the alignment behavior of single-species for $\theta \leq 1$. For $\theta = 1.5$, although the fluctuation is decaying in time, the tendency of alignment, $\delta\mathcal{E} \rightarrow 0$, seems unclear.

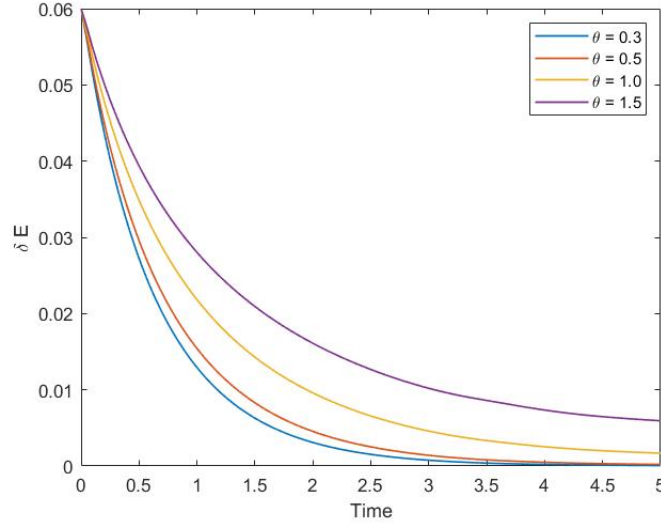


Figure 3.4.14: 1D single-species with fat-tail kernel, history of energy fluctuation up to $T = 5$. $N = 100$, $CFL = 0.7$.

3.4.3.2 Local communication

We also mention an interesting class of local communication kernels,

$$\phi(\mathbf{x}, \mathbf{x}') = \mathbf{1}_{\{d \leq |\mathbf{x} - \mathbf{x}'| \leq D\}}.$$

In this case, the communication between each two agents is activated only if their distance is within the range $[d, D]$. We consider the one-dimensional initial data (3.4.1) and keep track of the change in energy fluctuation, as shown in Figure 3.4.16. It seems that the energy fluctuation recovers polynomial decay under different choices of parameters. Also, a larger range of communication seems to result in a stronger tendency of alignment.

So far the effects of local communications have not been rigorously proved. Nevertheless, from the numerical experiments, we might expect that local communications can also lead to alignment, at least when the initial data is properly chosen, or if the range of communication is sufficiently large but remains finite. We will leave the theoretical discussion for future study.

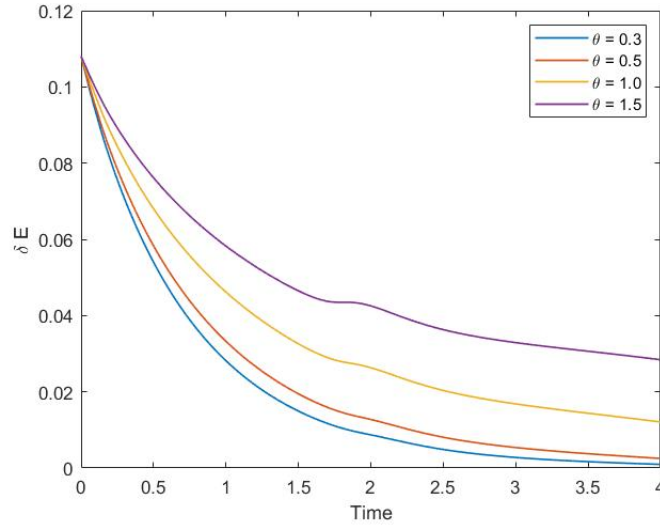


Figure 3.4.15: 2D single-species with fat-tail kernel, history of energy fluctuation up to $T = 4$. $N_x = N_y = 40$, $CFL = 0.7$.

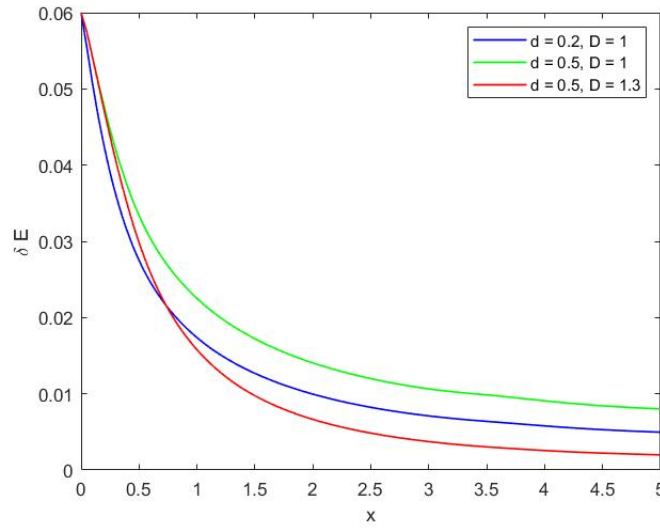


Figure 3.4.16: 1D single-species with local communication, history energy fluctuation up to $T = 5$. $N = 100$, $CFL = 0.7$.

Chapter 4: Novel Swarm-Based Gradient Descent Method

Besides successful application in mathematical modeling, the idea of swarm dynamics has also been widely applied in numerical optimization algorithms for non-convex problems. While local minimization problems have been well understood, finding the global minimum for a non-convex objective function is still a big challenge. Classical optimization algorithms, e.g. gradient descent method and Newton’s method, apply a single agent to explore the domain, with the marching direction determined from gradient information. Due to the limited local information, these methods often suffer from getting trapped near local minima and therefore are not suitable for global optimization problems. To help the iterations escape from local minima, the strategy of *swarm-based algorithms* is to employ multiple interacting agents (particles) to search for solution. With properly constructed interactions, the crowd is expected to gather around a dynamically established consensus state, which would give a near-optimal solution. Well-known examples include particle swarm optimization, ant colony optimization, genetic algorithm etc.

In this chapter, we propose a novel Swarm-Based Gradient Descent (SBGD) method for non-convex optimization. Discussions are based on the joint paper with Eitan Tadmor and Anil Zenginoglu [96]. In contrast to the other swarm-based methods where agents are treated ‘identically’, a central idea of SBGD is to distinguish the search strategies of different candidates based on their qualities. This is achieved by identifying each agent not only by its position, $\mathbf{x}$, but also a mass, m . The key to SBGD dynamics is communication: masses are being transferred from agents at high ground to low(-est) ground. At the same time, agents update positions with step size, $h = h(\mathbf{x}, m)$, which is dependent on their masses: heavier agents proceed with small time steps in the direction of local gradient, while lighter agents take larger time steps to escape from

local minima but still ensure the descent property. Under such mechanism, the crowd of agents is dynamically divided between ‘heavier’ leaders, which are expected to approach local minima, and ‘lighter’ explorers. With large-step protocol, light explorers are expected to encounter improved position for the swarm; if they do, then they will take over the role of ‘heavy’ swarm leaders. Convergence analysis and numerical simulations in one-, two-, and 20-dimensional benchmarks demonstrate the effectiveness of SBGD as a global optimizer.

4.1 The Swarm-Based Gradient Descent (SBGD) algorithm

Consider unconstrained global optimization problems of the form

$$\mathbf{x}_* = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}). \quad (4.1.1)$$

The Gradient Descent (GD) method is one of the most popular optimization algorithms and has been widely applied in a variety of application problems. It explores the surrounding area by marching along the direction dictated by local gradient, $\nabla f(\mathbf{x})$. Once the marching direction is determined, the choice of step size becomes the remaining key aspect, common protocols include fixed step size, exact line search and inexact line search. However, in non-convex global optimization problems, GD methods suffer from the lack of perception of global landscape, and therefore often get trapped in the basin of attraction of local minima.

To enhance the flexibility of exploration, the SBGD algorithm employs a swarm of agents, instead of a single agent, to search for the solution, each of which explores the ambient space under GD protocol. Indeed, by having N agents in the swarm, the result is expected to ‘ N times better’. Meanwhile, agents would dynamically exchange information with each other such that the optimal candidate would progressively dominate the search of the crowd. To this end, we identify each agent by its time dependent position, $\mathbf{X}_i(t^n) \in \Omega \in \mathbb{R}^d$, and its weight, $m_i(t^n) \in [0, 1]$. Communications between agents are characterized by dynamical mass transitions from high to low, hence the agents with better quality (lower ground) would be favored. The time step

protocol is then adjusted accordingly such that heavier agents take smaller time steps, whereas lighter agents proceed with larger time steps and explore larger regions away from current basins of attraction. In this way, these light explorers are expected to encounter a ‘better’ minimizing ground. Then, these light explorers are gradually converted into heavier, global leaders of the swarm. The dynamic distinction between heavy leaders and light explorers enables the approach towards local minimizers, while keep searching for even better global minimizers simultaneously.

We summarize the key ingredients as follows:

- **Agents.** Each agent is identified by its position, $\mathbf{X}_i(t) \in \mathbb{R}^d$, and its mass, $m_i(t) \in (0, 1]$. The total mass is kept constant in time, $\sum_i m_i(t) = 1$.
- **Protocol for time step.** The position of each agent is dynamically adjusted by taking a time step h_i in the gradient direction, $\nabla f(\mathbf{X}_i(t))$

$$\frac{d}{dt}\mathbf{X}_i(t) = -h_i \nabla f(\mathbf{X}_i(t)).$$

The time step, h_i , depends on the position of the agent at, $\mathbf{X}_i(t)$, *and* on its relative mass, $\tilde{m}_i(t)$,

$$\tilde{m}_i(t) := \frac{m_i(t)}{m_+(t)}, \quad m_+(t) = \max_i m_i(t).$$

The precise protocol for choosing the step size, based on backtracking, is outlined in Section 4.2.2. A key aspect is choosing h_i as a decreasing function of the relative mass, $\tilde{m}_i$: ‘heavier’ agents move slower, while ‘lighter’ agents take larger time steps. An alternative point of view is to interpret the $\tilde{m}_i$ ’s as the probabilities of agents to identify global minimum: those with mass $m_i(t) \ll m_+(t)$ take large time steps to explore the region of interest, since their probability of identifying the global minimum at their current position, $\mathbf{X}_i(t)$, is low.

- **Communication.** Let $f_{\max}(t) = \max_j f(\mathbf{X}_j(t))$ and $f_{\min}(t) = \min_j f(\mathbf{X}_j(t))$ denote the maximal and respectively, minimal heights of the swarm at time t . The mass of each agent,

$m_i(t)$, is dynamically adjusted according to its *relative height*, $\eta_i(t)$,

$$\begin{cases} \frac{d}{dt}m_i(t) = -\phi_p(\eta_i(t))m_i(t), & i \neq i(t), \\ m_i(t) = 1 - \sum_{j \neq i(t)} m_j(t), & i = i(t), \end{cases}$$

$$i(t) := \underset{i}{\operatorname{argmin}} f(\mathbf{X}_i(t)), \quad \eta_i(t) := \frac{f(\mathbf{X}_i(t)) - f_{\min}(t)}{f_{\max}(t) - f_{\min}(t)}.$$

At each iteration, each non-optimal agent loses a fraction of its mass, $\phi_p(\eta_i(t)) \in (0, 1]$, which is transferred to the current best minimizer at $\mathbf{X}_{i(t)}$. Here we allow to adjust the mass transition, $\phi_p(\eta) = \eta^p$, using a user choice of a fine-tuning parameter $p > 0$, with the default choice $p = 1$. As the best minimizer becomes ‘heavier’, it will be more ‘cautious’, taking smaller time steps, while the other ‘lighter’ agents take larger time steps. As the lighter agents explore the surrounding with larger time steps, it will increase their likelihood to encounter a new neighborhood of a global minimum, which in turn may place one of them as the new heaviest global minimizer and so on. Observe that with a larger p , the mass transition with rate $\phi_p(\eta_i(t))$ would be slower.

The discrete time marching of SBGD is realized by agents positioned at $\mathbf{X}_i^{n+1} = \mathbf{X}_i(t^{n+1})$ with masses $m_i^{n+1} = m_i(t^{n+1})$ at discrete time steps $t^{n+1} = t^n + \Delta t$. We use the simple forward time discretization with time step $\Delta t = 1$, acting on all non-empty agents, $m_i^n > 0$,

$$\left\{ \begin{array}{l} m_i^{n+1} = m_i^n - \phi_p(\eta_i^n)m_i^n, \quad i \neq i_n \\ m_{i_n}^{n+1} = m_{i_n}^n + \sum_{i \neq i_n} \phi_p(\eta_i^n)m_i^n, \\ \\ m_+^{n+1} := \max_i m_i^{n+1} \\ \mathbf{X}_i^{n+1} = \mathbf{X}_i^n - h(\mathbf{X}_i^n, \lambda\psi_q(\tilde{m}_i^{n+1}))\nabla f(\mathbf{X}_i^n), \quad \tilde{m}_i^{n+1} = \frac{m_i^{n+1}}{m_+^{n+1}} \end{array} \right\} \quad i_n := \underset{i}{\operatorname{argmin}}_i f(\mathbf{X}_i^n) \quad m_i^n > 0. \quad (4.1.2)$$

Initially, the agents are placed at random positions, $\{\mathbf{X}_i^0\}$ with equi-distributed masses $\{m_i^0 = 1/N\}$. At each iteration, masses of agents are exchanged according to their relative heights, and the positions of agents are updated in the direction of the local gradient, with time step, $h_i^n = h(\mathbf{X}_i^n, \lambda\psi_q(\tilde{m}_i^{n+1}))$, depending on these relative mass. Note that the agent with the worst configuration, positioned at $\operatorname{argmax} f(\mathbf{X}_n^i)$, is eliminated from the computation; consequently, the size of the swarm decreases by one agent at a time, until it remains with the one heaviest agent. Our choice for the time-stepping protocol, $h(\mathbf{X}, \lambda\psi_q(\tilde{m}))$, is the *backtracking line search* outlined in Section 4.2.2, which is weighted by the relative masses, $\psi_q(\tilde{m}_i^{n+1})$ (again, here we allow fine-tuning the dependence on the relative mass, $\psi_q(\tilde{m}) = \tilde{m}^q$, based on a user choice of $q > 0$, with the default choice $q = 1$). The backtracking enforces a descent property for the SBGD iterations $\mathbf{X}_i^n$, and the parameter, $\lambda \in (0, 1)$, dictates how much the descent property holds in the sense that (4.2.4) is fulfilled. The communication is designed so that the total mass of the swarm gradually concentrates on the agents most likely to become the global minimizers, that is, the agents which will most likely to reach the global minimum of the region explored so far by the swarm. Such ‘heavy’ agents are assigned with relatively small step sizes, as they are suspected to be close to ‘good’ minimizers, hence their subsequent explorations should be sufficiently cautious. On the other hand, the ‘lighter’ agents should not be trapped in basins of attraction of local minimizers, so they proceed with larger step sizes, allowing them to explore a larger regions, during which they may encounter ‘better’ minimizers; then they may be gradually converted from ‘light explorers’ into ‘heavy leaders’ and so on.

We note the flexibility of the SBGD communication protocol, depending on fractional mass transition, $\phi_p(\eta_i^n)$, and the mass-dependent step size, $h(\mathbf{X}_i^n, \lambda\psi_q(\tilde{m}_i^{n+1}))$. Their detailed construction is outlined in 4.2. The algorithm (4.1.2) then forms a family of swarm-based methods, denoted SBGD_{pq} whenever we want to emphasize its dependence on the parameters (p, q) ; the default version corresponding to $(p, q) = (1, 1)$ is denoted simply by SBGD.

4.1.1 Why communication is important

Consider the particular scenario in which all agents are assigned with the same constant mass, i.e., $m_i^n \equiv 1/N$ and $\eta_i^n \equiv 0$, so that $\psi_q(\tilde{m}_i^n) \equiv 1$ yields

$$\mathbf{X}_i^{n+1} = \mathbf{X}_i^n - h(\mathbf{X}_i^n, \lambda) \nabla f(\mathbf{X}_i^n), \quad i = 1, 2, \dots, N. \quad (4.1.3)$$

In this case, there is no mass transition and the dynamics is reduced to a crowd of *non-communicating* agents. In particular, each agent makes its own decision to proceed with variable step size based on the backtracking protocol outlined in section 4.2.2 below. We refer to this as the backtracking GD, or GD(BT) method. There is also the primary version of Gradient Descent, denoted GD(h), which proceeds with a fixed step size $h(\mathbf{X}_i^n, \lambda) \equiv h$. In either case, we have N agents, exploring the region of interest independently of each other. Of course, if there are N such agents exploring the region of interest, the corresponding GD(h) and GD(BT) method are expected to be “ N times better” than their the single-agent versions. However, compared with the swarm of N communicating agents, we find that the SBGD dynamics has a superior *global* behavior. Specifically, the advantage of communication in SBGD dynamics becomes apparent in exploring larger regions for potential global minimum. This will be borne out in the numerical results presented in sections 4.5.1, 4.5.2 and 4.5.3. Here, we demonstrate the benefit of communication with a simple example of an objective function shown in Figure 4.1.1,

$$f(x) = e^{\sin(2x^2)} + \frac{1}{10} \left(x - \frac{\pi}{2}\right)^2. \quad (4.1.4)$$

The function admits multiple local minima, with a unique global minimum ($x^* \approx 1.5355$). We compare the performance of SBGD_{pq}, (4.1.2) vs. the non-communicating GD(BT) iterations, (4.1.3), the GD(h) iterations with a fixed step size h , and the Adam method with (initial) step size α , denoted Adam(α), [79]. We report on the results of SBGD₂₁ which seems to perform slightly better than the basic version SBGD₁₁, and both offer a more robust optimizer than all

other non-communicating methods.

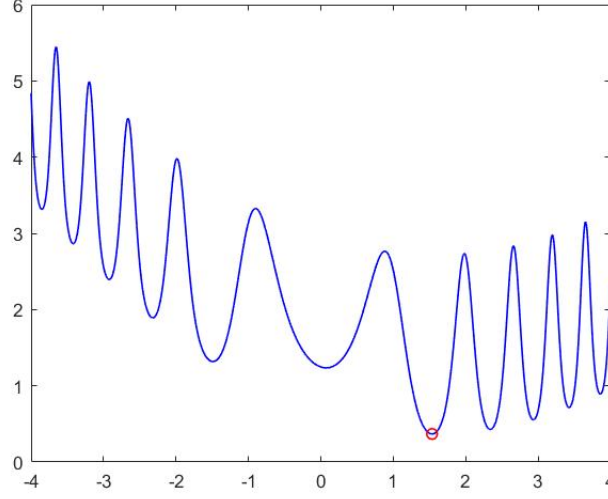


Figure 4.1.1: Plot of the objective function (4.1.4).

At first, we initialize the positions of agents uniformly in the interval $[-3, 3]$. In this case, the global minimum is included in the support of initial data. We implemented 1000 independent simulations and observe the results of SBGD, GD(h), GD(BT) and Adam. Table 4.1.1 presents the success rates for an increasing number of agents. All methods perform equally well in locating the global minimum, except for Adam(1.1): a large initial step size in Adam method may take it outside the initial region that already contains the global minimum.

The situation is different, however, if we initialize the agents to be uniformly distributed in $[-3, -1]$. The results shown in Table 4.1.2, indicate that the performance of the non-communicating GD(0.01), GD(BT) and Adam(0.1) is significantly worse, whereas the SBGD still identifies the global minimum with high success rates. In particular, the GD(h) and Adam with small time steps are trapped inside an initial basin of attraction, unable to get out of that neighborhood of local minimum. This is also depicted in Figure 4.1.2 where each local minimum sheds its local basin of attraction for GD(0.01) and Adam(0.1). In particular, the initial data outside $[1, 2]$ will necessarily fail to reach the global minimum at $x_* \approx 1.5355$. Only when combined with a larger initial step, Adam(1.1) leads to substantial improvement. Nevertheless, the behavior of Adam

method under large time step is more unstable. As observed from the computation of basin of attraction, a slight change in the initial position could lead to a significant jump in the resulting minimizer.

This is further clarified when we examine the distribution of $m = 1000$ solutions by SBGD vs. GD(BT) in Figure 4.1.3. Observe that in most of the 1000 experiments, the iterations of GD(BT) are blocked by the relatively flat basin near the origin, and subsequently they end at the local minimizer lying in the interval $[-2, -1]$. In contrast, the SBGD iterations, thanks to the ‘aggressive’ exploration of light agents, are much more likely to avoid getting trapped in the local flat basin of attraction and eventually accumulate enough mass nearby the global minimum.

N	5	10	15	20	30
SBGD ₁₁	64.3%	96.5%	99.8%	99.9%	100%
SBGD ₂₁	68.2%	97.7%	99.7%	100%	100%
GD(0.01)	75.2%	93.5%	98.7%	100%	100%
GD(BT)	73.6%	96.7%	99.5%	100%	100%
Adam(1.1)	19.6%	35.0%	64.9%	77.1%	89%
Adam(0.1)	58.3%	65.6%	85.8%	95.2%	95.7%

Table 4.1.1: Success rates of SBGD, GD(h), GD(BT) and Adam methods for global optimization of (4.1.4), based on $m = 1000$ runs with uniformly generated initial data in $[-3, 3]$. Backtracking parameters (see Algorithm 1), $\lambda = 0.2$ and $\gamma = 0.9$.

The example shows the advantage of communication-based SBGD as an algorithm that is more resilient to the initial guess. Specifically, in complicated applications it may not be realistic to ‘guess’ an initial configuration that encloses the unknown location of the global minimum, and consequently, GD(h), GD(BT) and Adam iterations may be trapped near a local minimizer due to ill-conceived initial guesses. In contrast, the final outcome of SBGD is more resilient with respect to the initial configuration, in exploring regions outside the enclosure of initial guesses. More can be found in numerical simulations recorded for one-, two- and 20-dimensional benchmark

N	5	10	15	20	30
SBGD ₁₁	36.5%	83.1%	97.2%	99.5%	100%
SBGD ₂₁	42.4%	91.4%	99.0%	99.8%	100%
GD(0.01)	0.0%	0.0%	0.0%	0.0%	0.0%
GD(BT)	1.8%	5.2%	8.5%	12.8%	21.8%
Adam(1.1)	40.2%	47.9%	82.7%	88.7%	93.9%
Adam(0.1)	0.0%	0.0%	0.0%	0.0%	0.0%

Table 4.1.2: Success rates of SBGD, GD(h), GD(BT) and Adam methods for global optimization of (4.1.4) based on $m = 1000$ runs of uniformly generated initial data in $[-3, -1]$. Backtracking parameters (see Algorithm 1), $\lambda = 0.2$ and $\gamma = 0.9$

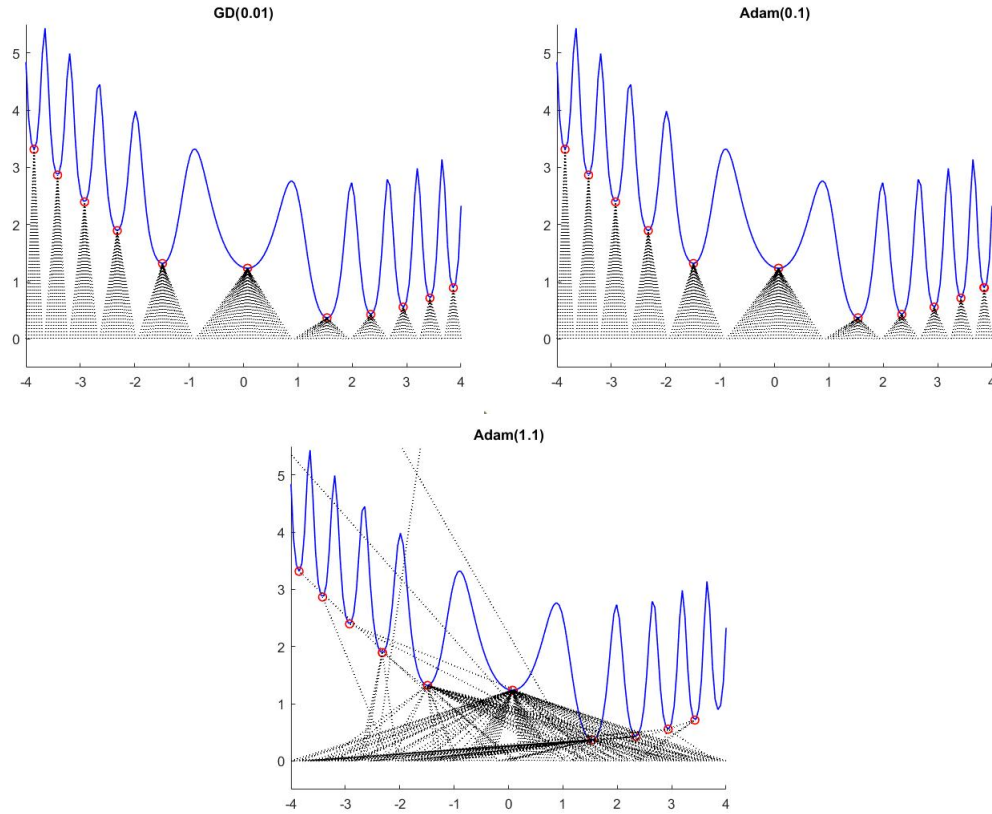
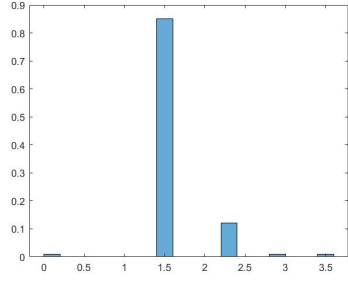
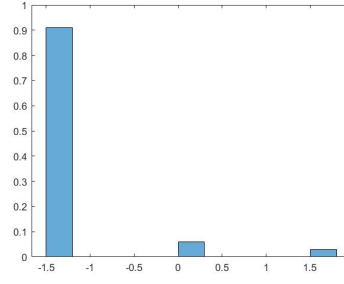


Figure 4.1.2: Basins of attraction for GD(0.01), Adam(0.1), and Adam(1.1) method.

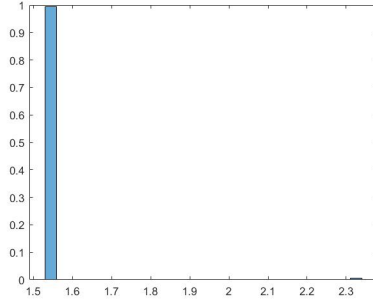
problems presented, respectively, in Sections 4.5.1, 4.5.2 and 4.5.3.



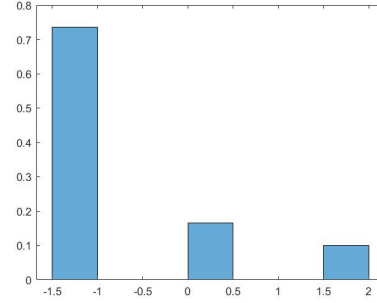
(a) SBGD, $N = 10$.



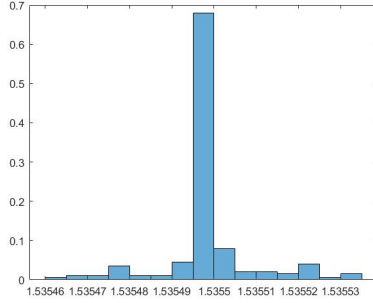
(b) GD(BT), $N = 10$.



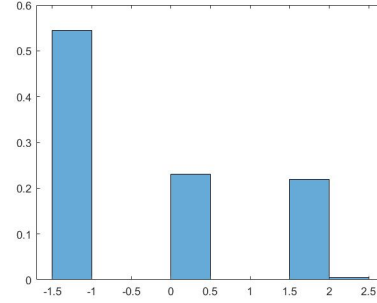
(c) SBGD, $N = 20$.



(d) GD(BT), $N = 20$.



(e) SBGD, $N = 30$.



(f) GD(BT), $N = 30$.

Figure 4.1.3: Histograms of problem (4.1.4) by $m = 200$ experiments. Initial data is generated uniformly in $[-3, -1]$. Global minimum at $x_* = 1.5355$.

4.1.2 What is different about SBGD

Besides our new SBGD method, there have also been many other swarm-based algorithms, notable examples include particle swarm optimization (PSO) [76], simulated annealing [80, 148], ant colony optimization [157], genetic algorithm [64], and the more recent *Consensus-Based*

Optimization (CBO) [18, 20, 116]. As a representative example, here we briefly mention the consensus-based optimization.

The CBO algorithm employs a crowd of interacting agents with positions $\mathbf{X}_1(t), \dots, \mathbf{X}_N(t) \in \mathbb{R}^d$ at time $t \in [0, \infty)$. Their motions are governed by the stochastic differential equations [20],

$$\begin{aligned} d\mathbf{X}_j(t) &= -\lambda(\mathbf{X}_j(t) - \mathbf{m}_N(t))dt + \sigma \sum_{k=1}^d (\mathbf{X}_j(t) - \mathbf{m}_N(t))_k dW_j^k(t) \vec{e}_k, \\ \mathbf{m}_N(t) &= \frac{1}{\sum_{j=1}^N e^{-\alpha f(\mathbf{X}_j(t))}} \sum_{j=1}^N \mathbf{X}_j(t) e^{-\alpha f(\mathbf{X}_j(t))}, \end{aligned} \quad (4.1.5)$$

subject to initial data $\mathbf{X}_j^0$. $\{W_j^k(t)\}$ are independent Brownian motions on $\mathbb{R}$, $\vec{e}_k$ is the unit vector along the k -th dimension. The weighted average state, $\mathbf{m}_N(t)$, serves as the center of attraction. By Laplace Principle [8], the exponential weights $e^{-\alpha f(\mathbf{X}_j)}$ concentrate the mass at the lowest agent, $\operatorname{argmin}_{\mathbf{X}_j} f(\mathbf{X}_j)$, for $\alpha \gg 1$. Hence, the drift enforces the *alignment* towards the *minimal heading*, while the stochastic diffusion is responsible for enlarging the range of exploration. It is shown that [20] with sufficiently large $\alpha \gg 1$ and appropriate constraints on parameters λ and σ , the crowd of agents will gradually gather around a consensus state, $\tilde{\mathbf{x}}$, lying in a neighborhood of global minimum $\mathbf{x}_*$ with radius $O(\frac{d \log(\alpha)}{2\alpha})$.

We point out that SBGD and CBO, to some extent, have similar mechanism. Indeed, communications of SBGD are characterized by mass transitions from agents at high ground to those at low ground. This can also be interpreted as a particular case of *alignment dynamics* (in terms of masses) where the agents steer towards the minimal heading, instead of steering towards the *average* heading as in the Cucker-Smale dynamics (see Chapter 2). Nevertheless, there are substantial differences between SBGD and CBO in the following aspects.

Gradient-based vs Gradient-free To resolve the problem of getting trapped at local minima, the CBO algorithm (and most of the other swarm-based algorithms) avoids the use of gradient information. As the search of CBO is directly based on *function values*, it is, theoretically, expected to be advantageous in resolving complex landscapes. However, in actual computation, especially in

high-dimensional problems, the application of α -weighted Laplace principle, requiring $\alpha \gg 1$, can suffer from significant machine error. Thus, due to the difficulty of taking sufficiently large α , the convergence of CBO is sensitive to the experimental choice of parameters λ and σ . In contrast, the search of SBGD is still based on function gradient. Although the theoretical optimality is not ensured, the quality of solution can be effectively improved by introducing proper communications. Moreover, the convergence of SBGD is *always* guaranteed regardless of the choice of artificial parameters (see Proposition 4.4.2).

Distinguished agents vs Identical agents In CBO algorithms, all agents are driven by ‘identical’ dynamics, taking the same formulation and uniform parameters. However, the central idea of SBGD is to distinguish different agents by identifying their qualities with masses m_i^n . This allows adaptive adjustments on individual search modes, in contrast to CBO of which the search parameters are fixed throughout the iterations. With dynamical and self-adjusting distinction between heavy leaders, which explore the surrounding more carefully, and light explorers, taking more aggressive step sizes to escape from current basins of attraction, the iterations would achieve a better balance between the stability of convergence and a large range of exploration. This is also verified by the numerical comparison of SBGD with the non-communicating GD and Adam methods.

4.2 Implementation of the SBGD_{pq} algorithm

The success of the SBGD_{pq} method relies heavily on two main procedures: (i) a properly defined communication protocol which dictates mass transition factors, $\{\phi_p(\eta_i^n)\}$; and (ii) an effective strategy for taking step size, $h(\mathbf{X}_i^n, \psi_q(\tilde{m}_i^{n+1}))$, which is adjusted to the position and the relative mass of a given agent. In this section, we discuss the details of these procedures, which are summarized in corresponding pseudo-codes.

4.2.1 Communications and mass transition

Let $i_n = \underset{i}{\operatorname{argmin}} f(\mathbf{X}_i^n)$ denote the global minimizer at iteration number n or time level t^n . All other agents will shed part (or all) of their mass, $\phi_p(\eta_i^n)m_i^n$, $i \neq i_n$, which will be transitioned to the mass of the global minimizer $m_{i_n}^n \mapsto m_{i_n}^n + \sum_{i \neq i_n} \phi_p(\eta_i^n)m_i^n$. The fraction of mass loss, η_i^n is determined by the *relative height*

$$\eta_i^n := \frac{f(\mathbf{X}_i^n) - f_{\min}^n}{f_{\max}^n - f_{\min}^n + \epsilon}, \quad (4.2.1)$$

where $f_{\min}^n = \min_j f(\mathbf{X}_j^n)$ and $f_{\max}^n = \max_j f(\mathbf{X}_j^n)$ are the current global extremes. To prevent vanishing denominator in the extreme case $f_{\max}^n = f_{\min}^n$, we introduce a small ϵ -correction, say $\epsilon = 10^{-10}$, the effect on the numerical performance is minimal. In this fashion, the ‘higher’ the agent is, the more mass it will lose. Particularly, for $\epsilon = 0$ the highest agent in each iteration will be eliminated. The function ϕ is user-dependent; for example $\phi_p(\eta) = \eta^p$, $p > 0$ enables adjusting the amount of mass transition, with mass transition tamed as $p \uparrow \infty$. This usage of a relative height keeps a minimal amount of global communication necessary to calibrate each agent relative to the current extremes of the crowd, while being invariant under the transition and dilation of the target function. Therefore, the computation of the relative height is more stable.

4.2.2 Time stepping – backtracking line search

Consider the Gradient Descent (GD) iteration

$$\mathbf{x}^{n+1} = \mathbf{x}^n - h \nabla f(\mathbf{x}^n).$$

The new position, $\mathbf{x}^{n+1} = \mathbf{x}^{n+1}(h)$, is viewed as a function of the step size, h . A proper strategy for choosing the step size is now the key for the success of GD iterations. We recall the classical *backtracking line search* (see e.g. [12]), which is a computational realization of the well-known Wolfe conditions [3, 155] for inexact line searches. The idea is to secure an acceptable step

length, h , that enforces a sufficient amount of height reduction (or loss) in the target function

$$f(\mathbf{x}^{n+1}(h)) \leq f(\mathbf{x}^n) - \lambda h |\nabla f(\mathbf{x}^n)|^2, \quad \lambda \in (0, 1). \quad (4.2.2)$$

Indeed, since $f(\mathbf{x}^{n+1}(h)) = f(\mathbf{x}^n) - h |\nabla f(\mathbf{x}^n)|^2 + \{\text{higher order terms}\}$, then (4.2.2) holds for any fixed $0 < \lambda < 1$, provided the time step is small enough, $h \ll 1$. The purpose is to secure (4.2.2) for large enough h , so that we maximize the size of descent, $\lambda h |\nabla f(\mathbf{x}^n)|^2$. To this end, one employs a dynamic adjustment, starting with a relatively large $h = h_0$ (for which one expects $f(\mathbf{x}^n - h \nabla f(\mathbf{x}^n)) > f(\mathbf{x}^n) - \lambda h |\nabla f(\mathbf{x}^n)|^2$) and then successively shrink the step size h until (4.2.2) is observed. At this stage, we have a final step size, $h = h(\mathbf{x}^n, \lambda)$, which secure (4.2.2)

$$f(\mathbf{x}^{n+1}(h)) \leq f(\mathbf{x}^n) - \lambda h(\mathbf{x}^n, \lambda) |\nabla f(\mathbf{x}^n)|^2.$$

Thus, the protocol for step size hinges on the choice of the parameter λ which dictates the steepness of descent (4.2.2). One expects that as the iterations approach the (potentially) global minimum, their descent property is scaled with a larger λ , which leads to a smaller step size, yet they should be able to avoid getting trapped in local basins of attraction by allowing smaller $\lambda \ll 1$. This is exactly where we take advantage of our swarm-based approach: as a compromise between these two conflicting requirements, we offer to use the relative mass of different agents, $\tilde{m}_i^{n+1}$, as an indicator which distinguishes between the ‘heavier’ agents who are potentially close to the global minimizer, and the ‘lighter’ agents which are allowed to take large(r) time steps. We therefore adjust the descent parameter λ to each SBGD_{pq} agent positioned at $\mathbf{X}_i^n$, according to its relative mass $\tilde{m}_i^{n+1}$

$$h_i^n = h(\mathbf{X}_i^n, \lambda \psi_q(\tilde{m}_i^{n+1})) \quad \tilde{m}_i^{n+1} := \frac{m_i^{n+1}}{m_+^{n+1}}, \quad m_+^{n+1} := \max_i m_i^{n+1}, \quad 0 < \lambda < 1. \quad (4.2.3)$$

The parameter $q > 0$ tunes the role of the relative mass $\tilde{m}_i^{n+1}$: as q increases then $\psi_q(\tilde{m}_i^{n+1})$ decreases, and backtracking allows intermediate agents to take larger time steps. The pseudo-

code for computing the SBGD_{pq} steps based on backtracking line search is given in Algorithm 1. The results reported in Section 4.5.2.2 below show that although fine-tuning the parameter, q , can lead to improved results, it has a limited effect on the overall performance of SBGD_{pq} iterations.

Algorithm 1 Backtracking Line Search

Set the descent parameter, $\lambda \in (0, 1)$, and shrinkage parameter, $\gamma \in (0, 1)$

Set ψ_q with $q > 0$

Set the relative mass $\tilde{m}_i^{n+1} = \frac{m_i^{n+1}}{m_+^{n+1}}$

Initialize the step size $h = h_0$.

while $f(\mathbf{X}_i^n - h\nabla f(\mathbf{X}_i^n)) > f(\mathbf{X}_i^n) - \lambda\psi_q(\tilde{m}_i^{n+1})h|\nabla f(\mathbf{X}_i^n)|^2$ **do**

$h \leftarrow \gamma h$.

end while

Set $h(\mathbf{X}_i^n, \lambda\psi_q(\tilde{m}_i^{n+1})) \leftarrow h$

Observe that the successive shrinking of step size in backtracking involves a shrinking factor $\gamma \in (0, 1)$: a small γ corresponds to a ‘crude’ line search while $\gamma \sim 1$, corresponds to a more refined search. Again, although fine-tuning the shrinkage parameter γ may lead to improved results, it does not seem to have a substantial effect on the overall performance of SBGD_{pq} .

The final output of Algorithm 1 yields, for each agent, an adjusted step size, $h_i^n = h(\mathbf{X}_i^n, \lambda\psi_q(\tilde{m}_i^{n+1}))$, which secures the descent property (4.2.2) with $\lambda\psi_q(\tilde{m}_i^{n+1})$ substituted for λ ,

$$f(\mathbf{X}_i^{n+1}) \leq f(\mathbf{X}_i^n) - \lambda\psi_q(\tilde{m}_i^{n+1})h_i^n|\nabla f(\mathbf{X}_i^n)|^2, \quad \mathbf{X}_i^{n+1} = \mathbf{X}_i^n - h_i^n\nabla f(\mathbf{X}_i^n). \quad (4.2.4)$$

Moreover, as we shall see in Lemma 4.4.1 below, the step sizes, h_i^n , admit a lower bound in terms of the corresponding relative masses of the different agents. The scaling of the step size using relative masses encodes the communication of different agents, which is the key to the success of the SBGD_{pq} algorithm. Roughly speaking, we can distinguish between two types of agents. The SBGD_{pq} iterations are led by the heavier agents, $m_i^{n+1} \approx m_+^{n+1}$, which tend to recover a

maximal local descent rate of order λ , (4.2.2). On the other hand, there are the lighter agents where $m_i^{n+1} \ll m_+^{n+1}$, which are less driven by the steepness of their decent, and are therefore better equipped as explorers of large areal search for the global minimizer. In this way, the mass-dependent adjustment of step size captures both the descent property of the target function while allowing the lighter explorers to pull away from local minimizers.

4.2.3 SBGD_{pq} pseudocode

The pseudocode of the SBGD_{pq} method is given in Algorithm 2. The initial setup consists of N randomly distributed agents $\mathbf{X}_1^0, \dots, \mathbf{X}_N^0$, associated with masses $m_1^0, \dots, m_N^0$. At the beginning, all agents are assigned equal masses, $m_j^0 = 1/N$, $j = 1, \dots, N$. At each iteration step, the agent $\mathbf{X}_{i_n} = \operatorname{argmin}_{\mathbf{X}_i^n} f(\mathbf{X}_i^n)$ attains the minimal value, while the other agents transfer part of their masses to the current optimal minimizer $\mathbf{X}_{i_n}$. Then all the agents are updated with the gradient descent method using the step lengths obtained with (4.2.3).

To further improve efficiency, we use three tolerance factors:

- If the mass of an agent is lower than a minimal threshold *tolm*, then this agent will be eliminated and its remaining mass will be transferred to the optimal agent at $\mathbf{X}_{i_n}$.
- ‘Sticking particles’. Agents that are sufficiently close to each other below a threshold *tolmerge*, are merged into a new agent, and their masses are combined into the newly generated agent.
- The iterations stop when the minimizer’s descent in two consecutive iterations is below a minimal threshold *tolres*.

Unless otherwise specified, all simulations reported in this paper employ the same thresholds

$$tolm = 10^{-4}, \quad tolmerge = 10^{-3}, \quad tolres = 10^{-4}. \quad (4.2.5)$$

The optimal choices of these thresholds are experimental. With smaller thresholds, *tolm* and

tolmerge, the SBGD algorithm will explore a larger part of the ambient space ending with a better solution, at the expense of reduced efficiency. A balance between the quality of the solution and the computational cost should be explored.

Algorithm 2 Swarm-Based Gradient Descent

Set three tolerance parameters, *tolm*, *tolmerge* and *tolres*
Set the adjustment parameters $p, q > 0$
Initialization:
Set N — the number of agents
Set initial positions $\mathbf{X}_1^0, \dots, \mathbf{X}_N^0$ randomly generated under initial distribution ρ_0
Set initial mass $m_1^0 = \dots = m_N^0 = 1/N$
Set the optimal agent, $i_0 = \underset{i}{\operatorname{argmin}} f(\mathbf{X}_i^0)$
for $n = 0, 1, 2, \dots$ **do**
Set $f_{\min}^n = f(\mathbf{X}_{i_n}^n)$, $f_{\max}^n = \max_i f(\mathbf{X}_i^n)$
for $i = 1, \dots, N$ and $i \neq i_n$ **do** % Mass transitions
 if $m_i^n < 1/N * \text{tolm}$ **then**
 set $m_i^{n+1} = 0$
 reduce the # of active agents: $N \leftarrow N - 1$
 else $m_i^{n+1} = m_i^n - \phi_p(\eta_i^n) m_i^n$ where $\eta_i^n = \frac{f(\mathbf{X}_i^n) - f_{\min}^n}{f_{\max}^n - f_{\min}^n}$.
 end if
end for
 $m_{i_n}^{n+1} = 1 - \sum_{i \neq i_n} m_i^{n+1}$ % The mass of the overall crowd is conserved
Compute $m_+ = \max_i m_i^{n+1}$
for $i = 1, \dots, N$ **do** % Gradient descent
 Compute relative masses $\tilde{m}_i^{n+1} = \frac{m_i^{n+1}}{m_+}$
 Compute the step size $h = h(\mathbf{X}_i^n, \lambda \psi_q(\tilde{m}_i^{n+1}))$ according to Algorithm 1.
 March: $\mathbf{X}_i^{n+1} = \mathbf{X}_i^n - h \nabla f(\mathbf{X}_i^n)$.
end for
Merge the agents if their distance $< \text{tolmerge}$.
Set the new optimal agent $i_{n+1} = \underset{i}{\operatorname{argmin}} f(\mathbf{X}_i^{n+1})$.
Compute the residual $\text{res} = \|\mathbf{X}_{i_{n+1}}^{n+1} - \mathbf{X}_{i_n}^n\|_2$
if $\text{res} < \text{tolres}$ **then**
 $\mathbf{X}_{\text{SOL}} \leftarrow \mathbf{X}_{i_{n+1}}^{n+1}$
 break
end if
end for

4.3 A general outlook

We point out that there are many possible extensions that can be worked out in connection with the SBGD algorithm, leading to a large class of swarm-based optimizers (SBO) with better communication protocols. We mention three of them.

General gradient descent directions Our SBGD approach can be used with a more general set of gradient descent directions, $\mathbf{p}_i^n$

$$\mathbf{X}_i^{n+1} = \mathbf{X}_i^n + h(\mathbf{X}_i^n, \lambda\psi_q(\tilde{m}_i^{n+1}))\mathbf{p}_i^n, \quad \langle \mathbf{p}_i^n, \nabla f(\mathbf{X}_i^n) \rangle < 0.$$

The convergence results in Section 4.4, with appropriate adjustments, remain valid. Moreover, the variety in choice of directions, other than local gradients, may offer a better ‘covering’ of the ambient space $\Omega \subset \mathbb{R}^d$.

Swarm-based optimization — a general paradigm The general paradigm for our swarm-based optimization is realized by embedding the d -dimensional ambient space in $(\mathbf{x}, \tilde{m}) \in \mathbb{R}^d \times [0, 1]$; here $\tilde{m}$ is an additional parameter space of masses/weights (or probabilities, or ‘fractional population’, etc) which serves as a communication platform for the crowd of agents positioned at $\{\mathbf{x}_i\} \in \mathbb{R}^d$. In this context, one can combine such communication-based swarm iterations with any single agent time-marching protocol. As examples we refer to the recent adaptive GD method [109] and the references therein. In the present work we use the time-marching protocol of gradient-descent, based on backtracking search. Other time marching protocols can be used.

Survival of the fittest The communication in SBGD is designed so that in each iteration, the ‘worst’ agent, positioned at $\mathbf{X}_{i_+}^n := \underset{\mathbf{X}_i^n}{\operatorname{argmax}} f(\mathbf{X}_i^n)$, is eliminated, as it loses all of its mass ($\eta_{i_+}^n = 1$). This policy of ‘survival of the fittest’ implies that the number of N initial active agents decreases in each iteration until the SBGD remains with only a single, ‘heaviest’ agent, which proceeds by the GD(BT) protocol. In particular, this policy implies that for small swarms, say

$N \sim 10$, the performance of SBGD iterations is expected to be similar or only slightly better than GD(BT) iterations, as shown in the numerical simulations reported in sections 4.5.2 and 4.5.3. Alternatively, one can design a less restrictive evolutionary policy that will allow ‘worst’ agents below a certain threshold to survive. This will evolve a larger set of explorers for longer times, with a greater chance of exploring new and better minima unseen before. Our numerical experiments show that a balanced policy for the ‘fittest’ can indeed have a substantial effect on the final result, at the expense of increased computational time.

4.4 Convergence and error analysis

The study of convergence and error estimates for the SBGD method requires to quantify the behavior of f . Here we emphasize that the required smoothness properties of f are only sought in the region explored by the SBGD iterations. We assume that there exists a *bounded* region, $\Omega \ni \mathbf{X}_i^n$ for all agents. Since the SBGD allows light agents to explore the ambient space with large step size (starting with h_0), we do not have apriori bound on Ω ; in particular, the footprint of the SBGD crowd $\text{conv}_i\{\mathbf{X}_i^n\}$ may expand well beyond its initial convex hull $\text{conv}_i\{\mathbf{X}_i^0\}$.

We consider the class of loss functions, $f \in C^2(\Omega)$, with Lipschitz bound on gradient

$$|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}|, \quad \forall \mathbf{x}, \mathbf{y} \in \Omega. \quad (4.4.1)$$

We begin by recalling the lower bound on the step size, secured by the backtracking line search in Algorithm 1.

Lemma 4.4.1 *Consider the SBGD_{pq} iterations (4.1.2), with step size $h_i^n = h(\mathbf{X}_i^n, \lambda\psi_q(\tilde{m}_i^{n+1}))$ determined by backtracking line search in Algorithm 1 with shrinkage factor $\gamma \in (0, 1)$. Then we have a lower bound on the step length*

$$h_i^n \geq \frac{2\gamma}{L}(1 - \lambda\psi_q(\tilde{m}_i^{n+1})). \quad (4.4.2)$$

Proof. By the Lipschitz continuity of ∇f ,

$$\begin{aligned} f(\mathbf{X}_i^n - h_i^n \nabla f(\mathbf{X}_i^n)) &\leq f(\mathbf{X}_i^n) - h_i^n |\nabla f(\mathbf{X}_i^n)|^2 + \frac{L}{2} (h_i^n)^2 |\nabla f(\mathbf{X}_i^n)|^2 \\ &= f(\mathbf{X}_i^n) - \left(1 - \frac{L}{2} h_i^n\right) h_i^n |\nabla f(\mathbf{X}_i^n)|^2, \end{aligned}$$

and hence, if h_i^n is small enough

$$h_i^n \leq \frac{2}{L} (1 - \lambda \psi_q(\tilde{m}_i^{n+1})) \rightsquigarrow f(\mathbf{X}_i^n - h_i^n \nabla f(\mathbf{X}_i^n)) \leq f(\mathbf{X}_i^n) - \lambda \psi_q(\tilde{m}_i^{n+1}) h_i^n |\nabla f(\mathbf{X}_i^n)|^2.$$

The backtracking line search iterations tell us that the inequality on the right holds for h_i^n but not for $\frac{h_i^n}{\gamma}$, and in particular, therefore, $\frac{h_i^n}{\gamma}$ must satisfy the reverse inequality on the left, that is, (4.4.2) holds. $\square$

4.4.1 Convergence to a band of local minima

Our next proposition provides a rather precise quantitative description for the convergence of the SBGD method. The convergence is determined by the time series of SBGD minimizers, $\{\mathbf{X}_-^n\}$,

$$\mathbf{X}_-^n = \mathbf{X}_{i_n}^n, \quad i_n = \operatorname{argmin}_i f(\mathbf{X}_i^n),$$

and the time series of its heaviest agents, $\{\mathbf{X}_+^n\}$,

$$\mathbf{X}_+^n = \mathbf{X}_{i_n^+}^n, \quad i_n^+ := \operatorname{argmax}_i m_i^n.$$

The interplay between minimizers and communication of masses leads to a gradual shift of mass, from higher ground to the minimizers. Eventually, when the SBGD minimizers gain enough mass to assume the role of heaviest agents, the two sequences coincide. Convergence is independent of the lighter agents.

We introduce the scaling $M = \max_i f(\mathbf{X}_i^0) - f(\mathbf{x}_*)$; since $f(\mathbf{X}_i^n)$ are decreasing, we

conclude that the SBGD iterations remain within that range, namely

$$\forall n, i : \max_{i,j} f(\mathbf{X}_j^n) - f(\mathbf{X}_i^n) \leq M, \quad M := \max_i f(\mathbf{X}_i^0) - f(\mathbf{x}_*) \quad (4.4.3)$$

To simplify matters we restrict our attention to the basic version of SBGD, $(p, q) = (1, 1)$.

Proposition 4.4.2 *Fix $\lambda \in (0, 1)$ and consider the SBGD iterations (4.1.2) with step size $h_i^n = h(\mathbf{X}_i^n, \lambda \tilde{m}_i^{n+1})$ determined by backtracking line search of algorithm 1 in (4.2.3). Let $\{\mathbf{X}_-^n\}_{n \geq 0}$ and $\{\mathbf{X}_{i_{n+1}}^n\}_{n \geq 0}$ denote the time sequence of SBGD minimizers and, respectively, parents of heaviest agents, at t^n . Then, there exists a constant, $C = C(\gamma, L, M, \lambda)$ given in (4.4.10) below, such that*

$$\sum_{n=0}^{\infty} \min \{ |\nabla f(\mathbf{X}_-^n)|, |\nabla f(\mathbf{X}_{i_{n+1}}^n)|, |\nabla f(\mathbf{X}_{i_{n+1}}^n)| \cdot |\nabla f(\mathbf{X}_-^n)| \}^2 < C \min_i f(\mathbf{X}_i^0). \quad (4.4.4)$$

Proof. By Lemma 4.4.1 we have $h_i^n \geq \frac{2\gamma}{L}(1 - \lambda \tilde{m}_i^{n+1})$ and the descent property (4.2.4) implies

$$\begin{aligned} f(\mathbf{X}_i^{n+1}) &\leq f(\mathbf{X}_i^n) - \lambda \tilde{m}_i^{n+1} h_i^n |\nabla f(\mathbf{X}_i^n)|^2 \\ &\leq f(\mathbf{X}_i^n) - \frac{2\gamma}{L}(1 - \lambda \tilde{m}_i^{n+1}) \lambda \tilde{m}_i^{n+1} |\nabla f(\mathbf{X}_i^n)|^2. \end{aligned} \quad (4.4.5)$$

The last descent bound applies to all agents, and we shall focus on that bound for the minimizer indexed with $i_n = \operatorname{argmin}_i f(\mathbf{X}_i^n)$. We distinguish between two scenarios.

The first scenario is the canonical scenario, in which the minimizer $i = i_n$ coincides with the heaviest agent, $m_{i_n}^{n+1} = m_+^{n+1}$; thus $\tilde{m}_i^{n+1} = 1$, and we conclude

$$f(\mathbf{X}_-^{n+1}) \leq f(\mathbf{X}_{i_n}^{n+1}) \leq f(\mathbf{X}_-^n) - \frac{2\gamma}{L}(1 - \lambda) \lambda |\nabla f(\mathbf{X}_-^n)|^2, \quad h_{i_n}^n \geq \frac{2\gamma}{L}(1 - \lambda). \quad (4.4.6)$$

The inequality on the left follows since $\mathbf{X}_-^{n+1} = \mathbf{X}_{i_{n+1}}^{n+1}$ is the global minimizer at t^{n+1} ; the inequality on the right follows from (4.4.5) with $i = i_n$.

A second scenario occurs when $m_{i_n}^{n+1} < m_+^{n+1}$, that is — when the mass of the minimizer

$m_{i_n}^{n+1}$ did not yet ‘catch-up’ with the heaviest agent. This means that there was a *parent* of heaviest agent positioned at, say $\mathbf{X}_{i_{n+1}}^n$, so that even after shedding a portion of its mass,

$$m_+^{n+1} = (1 - \eta_+^n) m_{i_{n+1}}^n,$$

it is still heavier than the minimizer at $\mathbf{X}_{i_n}^{n+1}$. Nevertheless, that minimizer *gained* the mass lost by the heaviest agent, and therefore, the *relative* mass of that minimizer is at least as large as

$$m_{i_n}^{n+1} > \eta_+^n m_{i_{n+1}}^n = \frac{\eta_+^n}{1 - \eta_+^n} m_+^{n+1} \rightsquigarrow \tilde{m}_{i_n}^{n+1} > \frac{\eta_+^n}{1 - \eta_+^n}.$$

Recall that the transition factors, η_i^n , are determined by the relative heights, (4.2.1), we conclude

$$\tilde{m}_{i_n}^{n+1} > \frac{f(\mathbf{X}_{i_{n+1}}^n) - f(\mathbf{X}_{i_n}^n)}{\max_j f(\mathbf{X}_j^n) - f(\mathbf{X}_{i_{n+1}}^n)} > \frac{1}{M} \left(f(\mathbf{X}_{i_{n+1}}^n) - f(\mathbf{X}_-^n) \right), \quad \mathbf{X}_-^n = \mathbf{X}_{i_n}^n.$$

We now have to consider two subcases of the second scenario, depending on the size of $f(\mathbf{X}_{i_{n+1}}^n) - f(\mathbf{X}_-^n)$.

Case (i). Assume $f(\mathbf{X}_{i_{n+1}}^n) - f(\mathbf{X}_-^n) \leq \frac{\gamma}{L} (1 - \lambda) \lambda |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2$. Appealing to the descent property (4.4.5) for the heaviest agent at $\mathbf{X}_+^{n+1}$, where $\tilde{m}_i^{n+1} = \tilde{m}_+^{n+1} = 1$, we find

$$\begin{aligned} f(\mathbf{X}_-^{n+1}) &\leq f(\mathbf{X}_+^{n+1}) \leq f(\mathbf{X}_{i_{n+1}}^n) - \frac{2\gamma}{L} (1 - \lambda) \lambda |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2 \\ &\leq f(\mathbf{X}_-^n) - \frac{\gamma}{L} (1 - \lambda) \lambda |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2. \end{aligned} \tag{4.4.7}$$

The inequality on the left follows since $\mathbf{X}_-^{n+1}$ is the global minimizer at t^{n+1} ; the middle inequality follows from the descent property (4.4.5) for the heaviest agent at $\mathbf{X}_+^{n+1}$, and the last inequality follows from our assumption.

Case (ii). Finally, we remain with the case

$$f(\mathbf{X}_{i_{n+1}}^n) - f(\mathbf{X}_-^n) \geq \frac{\gamma}{L} (1 - \lambda) \lambda |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2$$

which implies

$$\tilde{m}_{i_n}^{n+1} > \frac{1}{M} \left(f(\mathbf{X}_{i_{n+1}}^n) - f(\mathbf{X}_-^n) \right) \geq \frac{\gamma}{ML} (1 - \lambda) \lambda |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2.$$

As before, the descent property (4.4.5), together with the lower bound we secured for $\tilde{m}_{i_n}^{n+1}$ in this case, imply¹

$$\begin{aligned} f(\mathbf{X}_-^{n+1}) &\leq f(\mathbf{X}_{i_n}^{n+1}) \leq f(\mathbf{X}_{i_n}^n) - \frac{2\gamma}{L} (1 - \lambda \tilde{m}_{i_n}^{n+1}) \lambda \tilde{m}_{i_n}^{n+1} |\nabla f(\mathbf{X}_{i_n}^n)|^2 \\ &\leq f(\mathbf{X}_-^n) - \frac{\gamma^2}{ML^2} (1 - \lambda) \lambda^2 |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2 \cdot |\nabla f(\mathbf{X}_-^n)|^2. \end{aligned} \quad (4.4.8)$$

Combining (4.4.6), (4.4.7) and (4.4.8), we find

$$f(\mathbf{X}_-^{n+1}) \leq f(\mathbf{X}_-^n) - \frac{1}{C} \min \{ |\nabla f(\mathbf{X}_-^n)|^2, |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2, |\nabla f(\mathbf{X}_{i_{n+1}}^n)|^2 \cdot |\nabla f(\mathbf{X}_-^n)|^2 \}, \quad (4.4.9)$$

with

$$C = \max \left\{ \frac{L}{\gamma(1 - \lambda)\lambda}, \frac{ML^2}{\gamma^2(1 - \lambda)\lambda^2} \right\}, \quad (4.4.10)$$

and the desired bound (4.4.4) follows by a telescoping sum. $\square$

Remark 4.4.3 We observe that the summability bound (4.4.4) is driven by a worst case scenario alluded in case (ii) above. In this case, there is a potentially large difference of heights between the minimizing agent and heaviest agent; consequently, the descent property of the relatively lighter minimizer could be small and we had to rely on the descent property of the heaviest agent in (4.4.7), which led to (4.4.8).

The descent rates of different agents can be arbitrarily slow, due to their time-dependent mass. The summability bound (4.4.4) depends solely on the time sequence of SBGD minimizers, $\{\mathbf{X}_-^n\}$, and heaviest agents, $\{\mathbf{X}_+^n\}$, but it is independent of the lightweight agents. Eventually, for large enough n , the minimizers and heaviest agents of SBGD_{pq} coincide into one time sequence,

¹We can assume without loss of generality that $1 - \lambda \tilde{m}_{i_n}^{n+1} > 1/2$. Otherwise, the descent property is satisfied for any smaller λ .

$\{\mathbf{X}^n\}$. The key point is that time sub-sequences, $\{\mathbf{X}^{n_\alpha}\}$, satisfy a Palais-Smale condition [134, §II.2]: by monotonicity, $f(\mathbf{X}^{n_\alpha}) \leq \max_i f(\mathbf{X}_i^0)$ while $\nabla f(\mathbf{X}^{n_\alpha}) \xrightarrow{\alpha \rightarrow \infty} 0$.

At the end of SBGD iterations, we will end up with one or several minimizers taking the same height.

Theorem 4.4.4 *Consider the loss function $F \in C^2(\Omega)$ such that the Lip bound (4.4.1) holds and let $\{\mathbf{X}_-^n\}_{n \geq 0}$ denote the time sequence of SBGD minimizers, (4.1.2),(4.2.3). Then $\{\mathbf{X}_-^n\}_{n \geq 0}$ consists of one or more sub-sequences, $\{\mathbf{X}_-^{n_\alpha}, \alpha = 1, 2, \dots\}$, that converge to a band of local minima with equal heights,*

$$\mathbf{X}_-^{n_\alpha} \xrightarrow{n_\alpha \rightarrow \infty} \mathbf{X}_*^\alpha \text{ such that } \nabla f(\mathbf{X}_*^\alpha) = 0, \text{ and } f(\mathbf{X}_*^\alpha) = f(\mathbf{X}_*^\beta)$$

In particular, if f admits only distinct local minima in Ω (i.e., different local minima have different heights), then the whole sequence $\mathbf{X}_-^n$ converges to a minimum.

Proof. Since we assume the sequence $\{\mathbf{X}_-^n\}$ is bounded in Ω , it has a converging sub-sequences. Take *any* such converging sub-sequence $\mathbf{X}_-^{n_\alpha} \rightarrow \mathbf{X}_*^\alpha \in \Omega$. By (4.4.4), $\nabla f(\mathbf{X}_-^{n_\alpha}) \rightarrow 0$ for all sub-sequences, and hence $\mathbf{X}_*^\alpha$ are local minimizers, $\nabla f(\mathbf{X}_*^\alpha) = 0$. Moreover, since $f(\mathbf{X}_-^n)$ is decreasing and therefore converges to $\underline{f} = \inf_n f(\mathbf{X}_-^n)$, all $f(\mathbf{X}_*^\alpha)$ must have the same ‘height’, $\underline{f}$. The collection of equi-height minimizers $\{\mathbf{X}_*^\alpha \mid f(\mathbf{X}_*^\alpha) = f(\mathbf{X}_*^\beta)\}$ is the limit-set of $\{\mathbf{X}_-^n\}$. $\square$

4.4.2 Flatness and convergence rate

Theorem 4.4.4 indicates the convergence of SBGD without imposing any convexity condition on the loss function f , and therefore it comes without any rate. To quantify convergence *rate*, we need access to the fact that f should be ‘curved up’, at least within a sufficiently small neighborhood of a local minimum $\mathbf{X}_\alpha^*$. In the simplest case, f may be assumed to be locally convex. However, one must take into account that f may be more flat than just quadratic convexity. Indeed, these relatively flat local minima are the main hurdle in non-convex optimization. A precise classification for the level of ‘flatness’ is offered by the Lojasiewicz condition. According

Lojasiewicz inequality, [88, 94], if f is analytic in Ω then for every critical point of f , $\mathbf{x}_* \in \Omega$, there exists a neighborhood $\mathcal{N}_* \ni \mathbf{x}_*$ surrounding $\mathbf{x}_*$, an exponent $\beta \in (1, 2]$ and a constant $\mu > 0$ such that

$$\mu|f(\mathbf{x}) - f(\mathbf{x}_*)| \leq |\nabla f(\mathbf{x})|^\beta, \quad \forall \mathbf{x} \in \mathcal{N}_*. \quad (4.4.11)$$

The exponent β is associated to the *flatness* of ∇f at $\mathbf{x} = \mathbf{x}_*$: if $\nabla f(\mathbf{x})$ vanishes of order m at $\mathbf{x} = \mathbf{x}_*$, then $\beta = \frac{m+1}{m}$. In the particular case of local convexity, $\mathbf{x}_*$ is a simple minimum and (4.4.11) is reduced to the Poylack-Lojasiewicz condition [74, 117] corresponding to $\beta = 2$

$$\mu(f(\mathbf{x}) - f(\mathbf{x}_*)) \leq |\nabla f(\mathbf{x})|^2, \quad \forall \mathbf{x} \in \mathcal{N}(\mathbf{x}_*). \quad (4.4.12)$$

A smaller value of $\beta < 2$ indicates a more flat configuration of F in a region $\mathcal{N}_* \ni \mathbf{x}_*$.

In the theorem below, we restrict attention to SBGD_{1,1} and we assume that n is large enough which allows us to treat only the canonical scenario outlined in Proposition 4.4.2, where minimizers and heaviest agents coincide.

Theorem 4.4.5 *Consider the loss function $f \in C^2(\Omega)$ such that the Lip bound (4.4.1) holds, with minimal flatness β . Let $\{\mathbf{X}_-^n\}_{n \geq 0}$ denote the time sequence of SBGD minimizers, (4.1.2), (4.2.3). Then, there exists a constant, $C = C(\gamma, \lambda, \mu)$, such that*

$$f(\mathbf{X}_-^{n_\alpha}) - f(\mathbf{X}_*^\alpha) \begin{cases} \leq \left(1 - \frac{2\mu\gamma\lambda(1-\lambda)}{L}\right)^{n_\alpha} \left(\min_i f(\mathbf{X}_i^0) - f(\mathbf{x}_*)\right), & \beta = 2 \\ \lesssim C \left(\frac{1}{n_\alpha}\right)^{\frac{\beta}{2-\beta}}, & \beta \in (1, 2) \end{cases} \quad (4.4.13)$$

Remark 4.4.6 *Observe that as ‘flatness’, increases, β is decreasing and the exponential decay in (4.4.13)₁ is replaced by a polynomial decay which may slow down all the way to first-order decay, $1/n_\alpha$.*

Proof. As before, we start with the descent property for $\mathbf{X}_-^n = \mathbf{X}_{i_n}^n$. The backtracking for $\mathbf{X}_i^n$ for

$i = i_n$ secures the descent property (4.4.6)

$$f(\mathbf{X}_-^{n+1}) \leq f(\mathbf{X}_-^n) - \lambda h_- |\nabla f(\mathbf{X}_-^n)|^2, \quad h_{i_n}^n \geq h_- := \frac{2\gamma}{L}(1 - \lambda).$$

We focus on the converging sub-sequence $\{\mathbf{X}_-^{n_\alpha}\}$,

$$f(\mathbf{X}_-^{n_\alpha+1}) \leq f(\mathbf{X}_-^{n_\alpha}) - \lambda h_- |\nabla f(\mathbf{X}_-^{n_\alpha})|^2.$$

Let us first discuss the quadratic case, $\beta = 2$, of Polyak-Lojasiewicz condition (4.4.12), which yields

$$f(\mathbf{X}_-^{n_\alpha+1}) \leq f(\mathbf{X}_-^{n_\alpha}) - \mu \lambda h_- (f(\mathbf{X}_-^{n_\alpha}) - f(\mathbf{X}_*^\alpha)), \quad \mathbf{X}_-^{n_\alpha} \in \mathcal{N}_\alpha.$$

Rearranging we find

$$f(\mathbf{X}_-^{n_\alpha+1}) - f(\mathbf{X}_*^\alpha) \leq (1 - \mu \lambda h_-) (f(\mathbf{X}_-^{n_\alpha}) - f(\mathbf{X}_*^\alpha)), \quad (4.4.14)$$

which yields exponential rate, [74, 117]

$$f(\mathbf{X}_-^{n_\alpha}) - f(\mathbf{X}_*^\alpha) \leq (1 - \mu \lambda h_-)^{n_\alpha} (f(\mathbf{X}_-^0) - f(\mathbf{X}_*^\alpha)).$$

The case of general Lojasiewicz bound (4.4.11) with $\beta < 2$ yields that the error, $E_{n_\alpha} := f(\mathbf{X}_-^{n_\alpha}) - f(\mathbf{X}_*^\alpha)$, satisfies

$$E_{n_\alpha+1} \leq E_{n_\alpha} - \lambda h_- (\mu E_{n_\alpha})^{2/\beta}, \quad \mathbf{X}_-^{n_\alpha} \in \mathcal{N}_\alpha.$$

The solution of this Riccati inequality (e.g., [139, Theorem 3.1] for the limiting case $\beta = 1$), yields

$$f(\mathbf{X}_-^{n_\alpha}) - f(\mathbf{X}_*^\alpha) \lesssim_\mu \left\{ |\min_i f(\mathbf{X}_i^0) - f(\mathbf{X}_*^\alpha)|^{-1/\beta'} + \lambda h_- \mu^{2/\beta} n_\alpha \right\}^{-\beta'}, \quad \beta' = \frac{\beta}{2 - \beta} > 1.$$

and (4.4.13) follows with $C = \left(\frac{L}{2\gamma\lambda(1-\lambda)\mu^{2/\beta}} \right)^{\frac{\beta}{2-\beta}}$. □

4.5 Numerical results

4.5.1 One-dimensional problems

We use the swarm-based gradient descent method to search for the global minimizers of the 1D functions

$$\text{Ackley function : } f_{\text{Ackley}}(x) = -20e^{-0.2|x-B|} - e^{\cos(2\pi(x-B))} + 20 + e + C,$$

$$\text{Rastrigin function : } f_{\text{Rstrgn}}(x) = (x-B)^2 - 10\cos(2\pi(x-B)) + 10 + C,$$

with shift parameters B, C . Figure 4.5.1 shows that both, the Ackley and the Rastrigin function attain multiple local minimizers.

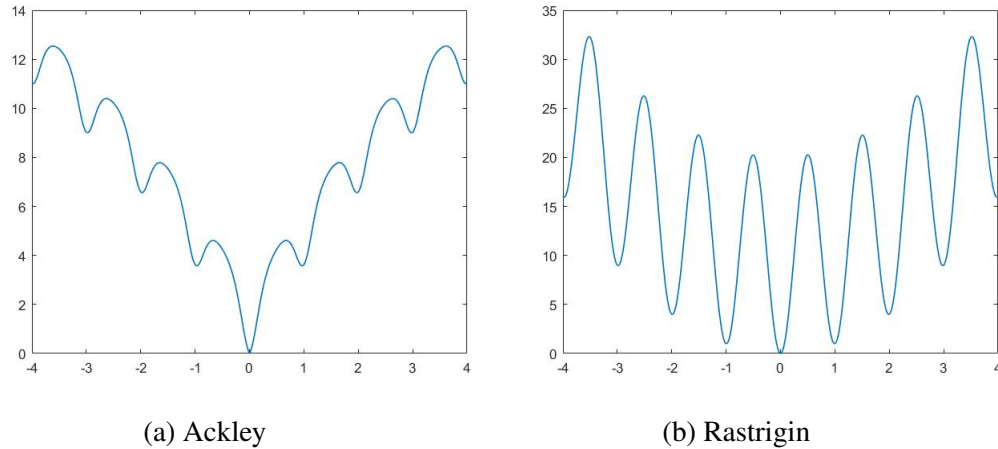


Figure 4.5.1: Benchmark functions

We implemented the ‘basic’ version of SBGD method $(p, q) = (1, 1)$ (i.e., $\phi_p(\eta) = \eta$ and $\psi_q(\tilde{m}) = \tilde{m}$), and with parameters $\lambda = 0.2$, $\gamma = 0.9$, $h_0 = 1$. Here λ, γ are the descent and shrinkage parameters, and h_0 is the initial step length for backtracking line search. We record the results of $m = 200$ independent simulations of Algorithm 2 with initial positions of the agents uniformly distributed in the interval $[-3, 3]$.

To illustrate the behavior of the algorithm, we present the evolution of agents in one representative simulation for the Ackley function and the Rastrigin function in Figures 4.5.2 and 4.5.3, where $N = 20$ agents are applied to search the domain. The blue line depicts the target function, the red circles represent the agents. It is seen that as the iterations progress, the agents flock towards different minima. The agents getting stuck at local minimizers are gradually removed due to the mass transition, while those approaching the global minimizer are eventually merged into one agent. Table 4.5.1 shows the results for the Ackley and the Rastrigin functions with a varying number of agents. The algorithm gives accurate solutions in all the test cases. With more agents applied the quality of the approximation is improved. However, we notice that improvement is significant as N increases from 5 to 10, whereas increasing the number of agents from 10 to 20 does not result in much enhancement.

B=0		N=5	N=10	N=20
Ackley function f_{Ackley}	success rate	99.50%	100%	100%
	$\mathbb{E} X_{\text{SOL}} - x_* ^2$	$4.70e^{-3}$	$8.21e^{-10}$	$6.89e^{-10}$
Rastrigin function f_{Rstrgn}	success rate	97.00%	100%	100%
	$\mathbb{E} X_{\text{SOL}} - x_* ^2$	$4.45e^{-2}$	$5.18e^{-10}$	$1.89e^{-10}$

Table 4.5.1: Results of SBGD for 1D Ackley and Rastrigin functions, $m = 200$.

The key feature of the SBGD algorithm is communication, reflected by the mass transitions between the agents. Such a mechanism dynamically adjusts the search strategies of different individuals. It is of interest to verify the benefit of communication. We compare the results by the SBGD method with those by the non-communicating multi-agents gradient descent, GD(BT), of which all the agents conduct gradient descent search independently. Tables 4.5.2 and 4.5.3 present the results by SBGD and GD(BT) methods obtained from $m = 200$ experiments computed with varying shift parameter B and different numbers of agents. We observe that when the initial distribution centers at $\mathbf{x}_*$ or is moderately shifted, both SBGD and GD(BT) methods are able to

find the global minimizer. However, the advantage of SBGD is more pronounced as $\mathbf{x}_*$ is shifted farther away from the center of initial data, and the non-communicating GD(BT) fails to give correct solutions since its exploration is restricted to the neighborhood of its initial data. Figure 4.5.4 displays the scenario $B = 25$, where the initial distribution is strongly shifted away from the global minimizer. The GD(BT) fails to find the global minimum. In contrast, the SBGD method employs light agents to conduct a more aggressive search in a larger area, and the algorithm ends at the correct minimizer at a surprisingly high success rate.

Remark 4.5.1 *Observe that when the initial data are placed far from the global minimum, the presence of a strong shift, $B = 25$, requires sufficiently many swarming agents $N > 20$, in order to secure a success rate $> 90\%$ and drive the expected value of the error, $\mathbb{E}|\mathbf{X}_{SOL} - \mathbf{x}_*|^2 < 0.5$.*

$\mathbf{x}_* = B$		N=10	N=20	N=30
$B = 0$	success rate	100%	100%	100%
	$\mathbb{E} X_{SOL} - x_* ^2$	8.42e-10	8.37e-10	3.38e-10
$B = 5$	success rate	100%	100%	100%
	$\mathbb{E} X_{SOL} - x_* ^2$	8.41e-10	7.58e-10	5.01e-10
$B = 15$	success rate	98.5%	100%	100%
	$\mathbb{E} X_{SOL} - x_* ^2$	1.41e-02	4.69e-03	8.27e-10
$B = 25$	success rate	45.5%	89.0%	98.5%
	$\mathbb{E} X_{SOL} - x_* ^2$	1.48e+02	1.49e+01	3.28e-01

Table 4.5.2: Shifted 1D Ackley, results by SBGD, $m = 200$.

4.5.2 Two-dimensional problems

Let us first comment on the general issue of optimization of parameter space. As noted in [154], once the parameterization of an adaptive GD method is fixed, it may not yield as good

$\mathbf{x}^* = B$		N=10	N=20	N=30
$B = 0$	success rate $\mathbb{E} X_{SOL} - x^* ^2$	100% 8.60e-10	100% 1.36e-09	100% 1.29e-09
$B = 5$	success rate $\mathbb{E} X_{SOL} - x^* ^2$	100% 8.51e-10	100% 1.25e-09	100% 1.21e-09
$B = 15$	success rate $\mathbb{E} X_{SOL} - x^* ^2$	46.5% 6.44e+01	75.0% 2.26e+01	85.5% 1.14e+01
$B = 25$	success rate $\mathbb{E} X_{SOL} - x^* ^2$	0% 4.86e+02	0% 4.50e+02	0% 4.38e+02

Table 4.5.3: Shifted 1D Ackley, results by GD(BT) (no communication), $m = 200$.

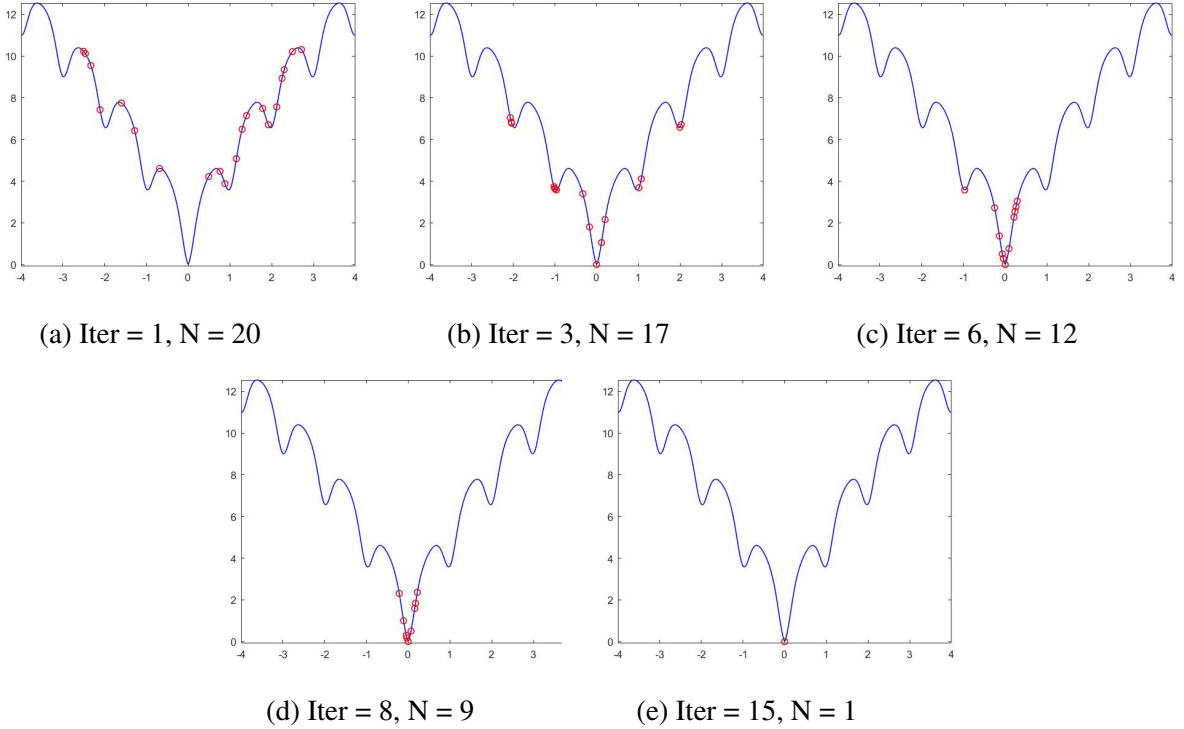


Figure 4.5.2: 1D Ackley: evolution of agents in one simulation. $B = C = 0$.

as or better results as simpler GD methods, yet adaptivity should be ‘judged’ after being optimized in space parameter, [25]. In this context, one can argue that optimizing a single agent method in parameter space is equivalent to a selective choice among many simulations of non-communicating multi-agent dynamics, whereas the swarm-based approach provides a *dynamic*,

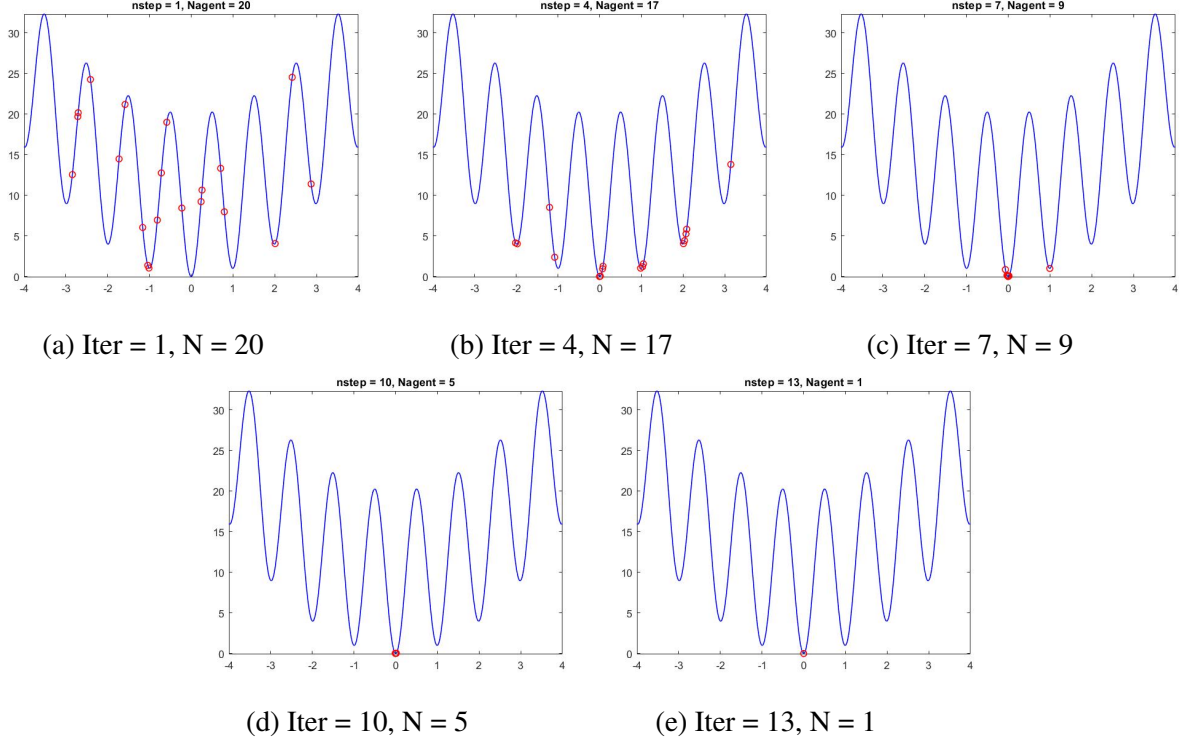
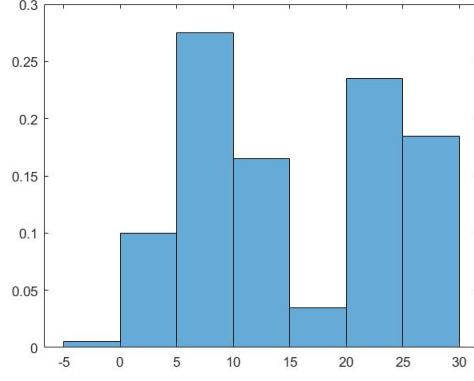


Figure 4.5.3: 1D Rastrigin: evolution of agents in one simulation. $B = C = 0$.

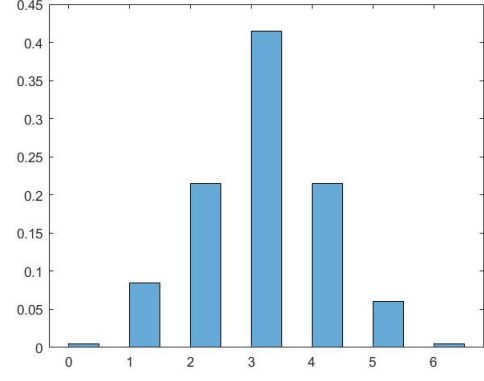
‘on the fly’ selection of optimized parameters, which is precisely the type of comparisons we make below.

On the other hand, our SBGD method depends on several parameters: the initial step h_0 , the descent parameter λ and shrinkage parameter γ tied to the backtracking, the tolerance parameters (4.2.5) and the (p, q) parameters. In the multi-dimensional computations reported below, we do *not* optimize these SBGD parameters. Thus, unless otherwise stated, we examine the performance of the SBGD method in Algorithm 2 with initial step size $h_0 = 1$, a descent parameter $\lambda = 0.2$, a shrinkage parameter $\gamma = 0.9$ and the threshold parameters (4.2.5). We begin here with the basic version of SBGD, $(p, q) = (1, 1)$, although later we shall find out that the choice $(p, q) = (2, 1)$ seems universally better. We run m number of independent simulations, initiated with uniformly distributed positions, $\{\mathbf{X}_i^0\}$ and measure the success of the SBGD according to the proportion of its successful end results.

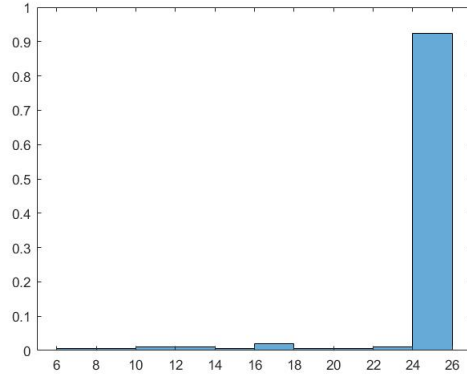
We illustrate the performance of the SBGD algorithm in multiple dimensions on three



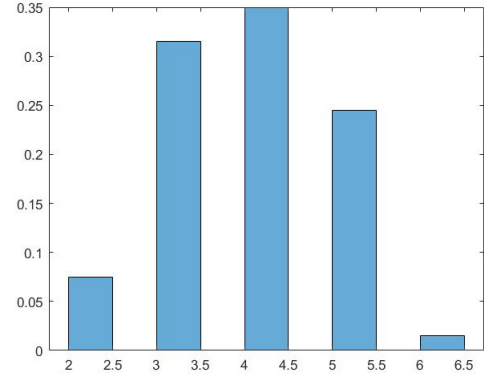
(a) SBGD, $N = 10$



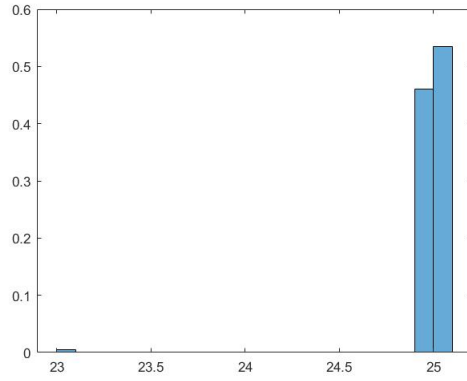
(b) GD(BT), $N = 10$



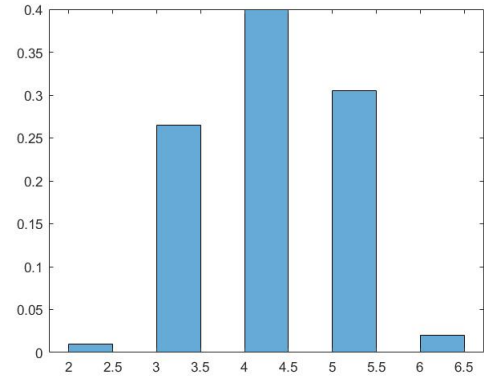
(c) SBGD, $N = 20$



(d) GD(BT), $N = 20$



(e) SBGD, $N = 20$



(f) GD(BT), $N = 20$

Figure 4.5.4: Histograms of the shifted 1D Ackley function by 200 experiments. $B = 25$, $C = 5$. Global minimum $\min f_{\text{Ackley}} = 5$ is attained at $\mathbf{x}_* = 25$.

benchmark test cases, [70]. First, the Ackley function

$$f_{\text{Ackley}}(\mathbf{x}) = -20 \exp \left\{ -\frac{0.2}{\sqrt{d}} \left\{ \sum_{i=1}^d (x_i - B)^2 \right\}^{1/2} \right\} - \exp \left\{ \frac{1}{d} \sum_{i=1}^d \cos(2\pi(x_i - B)) \right\} + 20 + e + C. \quad (4.5.1)$$

Second, the Rastrigin function

$$f_{\text{Rstgin}}(\mathbf{x}) = \frac{1}{d} \sum_{i=1}^d \left\{ (x_i - B)^2 - 10 \cos(2\pi(x_i - B)) + 10 \right\} + C. \quad (4.5.2)$$

Here d is the dimension of the ambient space, $\mathbf{x} := (x_1, \dots, x_d)$ is the variable in $\mathbb{R}^d$ and $B, C \in \mathbb{R}$ are the shift parameters. Both functions attain the global minimum C at the unique global minimizer $\mathbf{x}_* = B$.

Third, we consider the drop-wave function

$$f_{\text{Drop}}(\mathbf{x}) = -\frac{1 + \cos(12|\mathbf{x}|)}{0.5|\mathbf{x}|^2 + 2}, \quad (4.5.3)$$

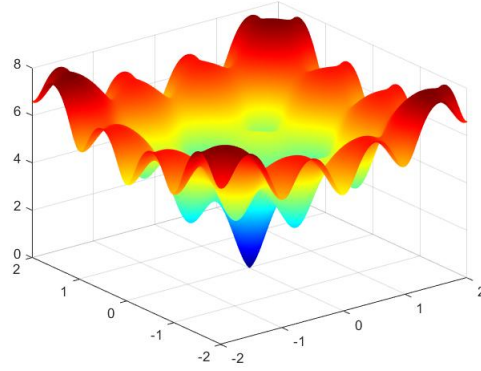
with SBGD parameters $\lambda = 0.3$, $\gamma = 0.9$, $h_0 = 1$. The agents are initialized with uniform distribution of positions in the hypercube $[-3, 3]^d$.

To evaluate the quality of the solution, we make use of the *success rate* among m independent simulations. We consider a simulation to be successful if $\mathbf{X}_{\text{SOL}}$ is within the d -dimensional cube $[\mathbf{x}_* - 0.25, \mathbf{x}_* + 0.25]^d$, e.g., [20, §4.2]. This condition ensures that the approximate solution lies in the basin of attraction of the global minimizer. In fact, in a successful experiment the solution will lie in a much smaller neighborhood of $\mathbf{x}_*$.

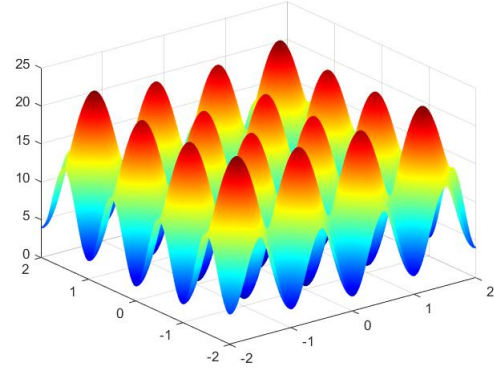
4.5.2.1 SBGD compared with non-communicating GD(BT)

We verify the advantage of the SBGD method in comparison to the non-communicating GD(BT) algorithm. We consider the three benchmarks of Ackley, Rastrigin and drop-wave functions in two dimensions. The landscapes are as shown in Figure 4.5.5. All functions have multiple local minimums. While the global minimum of the Ackley function is obviously lower than the other local minimums, the global minimum of the Rastrigin function is much less distinguishable. The drop-wave function has complex geometry with high frequency local minima and sharp basins of attractions. Its global minimum $f(\mathbf{x}_*) = -1$ is attained at the unique global minimizer

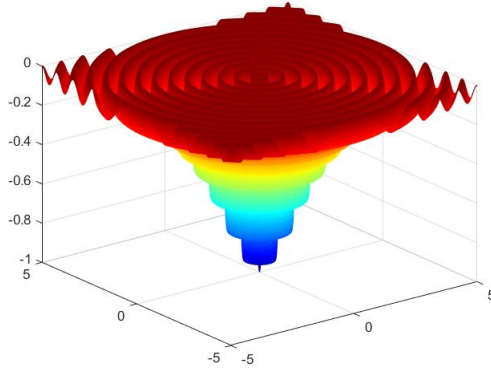
$$\mathbf{x}_* = 0.$$



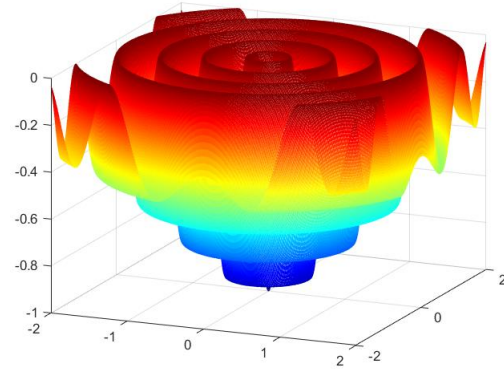
(a) 2-D Ackley



(b) 2-D Rastrigin



(c) 2-D drop-wave function



(d) zoom into 2-D drop-wave function

Figure 4.5.5: Landscapes of benchmark functions in two dimensions

Table 4.5.4 compares the success rates of SBGD and GD(BT) methods in the Ackley test cases with varying function shifts B and different numbers of agents N . It is observed that the two methods perform comparably well when the initial distribution is centered at $\mathbf{x}_*$ or moderately shifted. But as the shift parameter B increases, the SBGD method still achieves high success rates whereas the non-communicating GD(BT) method fails. The results for the Rastrigin function, given in Table 4.5.5, echo the same advantage of SBGD in shifted scenarios.

Table 4.5.6 compares the success rates of SBGD and GD(BT) methods in the test case of the 2D drop-wave function, (4.5.3). We compare the success rates of SBGD with $\lambda = 0.3$ vs. the non-communicating GD(BT) using $m = 200$ simulations randomly initiated with uniform distribution

$\mathbf{x}_* = B$		N=25	N=50	N=100
$B = 0$	SBGD	98.0%	100%	100%
	GD(BT)	100%	100%	100%
$B = 5$	SBGD	93.6%	98.6%	99.8%
	GD(BT)	71.2%	87.0%	99.2%
$B = 10$	SBGD	66.2%	90.8%	98.4%
	GD(BT)	0%	0%	0.6%

Table 4.5.4: 2D Ackley, success rates of SBGD vs. GD(BT), $m = 500$.

$\mathbf{x}_* = B$		N=25	N=50	N=100
$B = 0$	SBGD	73.6%	95.2%	100%
	GD(BT)	53.0%	76.4%	96.40%
$B = 5$	SBGD	44.4%	80.4%	99.2%
	GD(BT)	15.8%	28.4%	56.4%
$B = 10$	SBGD	14.8%	44.0%	78.4%
	GD(BT)	13.0%	30.4%	56.2%

Table 4.5.5: 2D Rastrigin, success rates of SBGD vs. GD(BT), $m = 500$.

at $[-3, 3]^2$. The SBGD results reported in Table 4.5.6 show a remarkable improvement over the performance of GD(BT); communication helps.

The SBGD does not always offer such a decisive lead over GD(BT). In table 4.5.7 we compare the error of SBGD vs. GD(BT) for the 2D Rosenbrock function

$$f_{\text{Rsnbrk}}(\mathbf{x}) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2. \quad (4.5.4)$$

	N=10	N=20	N=30
SBGD	90.5%	99.5%	100%
GD(BT)	15.0%	21.5%	35.5%

Table 4.5.6: 2D drop-wave, success rates of SBGD vs. GD(BT), $m = 500$.

$\mathbb{E} \mathbf{X}_{SOL} - \mathbf{x}_* ^2$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$
SBGD	2.07e-1	6.18e-2	1.29e-1
GD(BT)	4.25e-2	6.08e-2	1.29e-1

Table 4.5.7: Average error for 2D Rosenbrock function SBGD vs. GD(BT), $m = 500$ simulations equi-distributed at $[-4, -2]^2$.

The stiffness of f_{Rsnbrk} near its global minimum at $\mathbf{x}_* = (1, 1)$ produces comparable results of the swarm dynamics and non-communicating GD(BT).

4.5.2.2 SBGD_{pq} method – dependence of (p, q) parameters

We now turn to discuss the effect of adjusting the mass transition with $\phi_p(\eta_i^n)$ and back-tracking with $\psi_q(\tilde{m}_i^{n+1})$.

We implemented the SBGD_{pq} for shifted 2D Rastrigin function (4.5.2) with shift $B = 5$, and drop-wave function (4.5.3) using SBGD_{pq} with descent parameter $\lambda = 0.3$, using N agents uniformly initialized at $[-3, 3]^2$. The success rates among $m = 500$ simulations with different (p, q) are recorded in Tables 4.5.8 and 4.5.9. In both cases, the results for the basic SBGD, $(p, q) = (1, 1)$, with a success rate of 80% and respectively 90%, are substantially improved to 97%, once we use SBGD_{pq} with $(p, q) = (2, 1/2)$. In this case, the use of $p = 2$ enforces

more moderate mass transitions which seems to play a key role whenever f has steep basins of attractions, while $q = 1/2$ enforces a slower marching protocol for intermediate agents.

$\phi_p(\eta)\psi_q(\tilde{m})$	$\sqrt{\tilde{m}}$	$\tilde{m}$	$\tilde{m}^2$
$\sqrt{\eta}$	77.2%	43.0%	40.0%
η	94.8%	80.4%	73.4%
η^2	97.8%	94.0%	88.4%
η^{20}	83.4%	95.4%	98.2%

Table 4.5.8: SBGD_{pq} for shifted 2D Rastrigin function; $N = 50$ agents

$\phi_p(\eta)\psi_q(\tilde{m})$	$\sqrt{\tilde{m}}$	$\tilde{m}$	$\tilde{m}^2$
$\sqrt{\eta}$	93.6%	83.0%	79.8%
η	97.2%	90.4%	85.2%
η^2	97.6%	96.0%	90.2%
η^{20}	80.8%	98.8%	97.4%

Table 4.5.9: SBGD_{pq} for 2D Drop-wave function; $N = 10$ agents

As a further example we consider the 2D Rosenbrock function f_{Rsnbrk} in (4.5.4). The difficulty arises when the iterations approach the neighborhood of a global minimum $f(\mathbf{x}_*) = 0$ attained at $\mathbf{x}_* = (1, 1)$. This is due to the severe ‘skew-ness’ of f which is sensitive to the descent parameter λ . We employ $N = 30$ agents initialized with randomly distributed positions at $[-4, -2]^2$. Success rates are computed among $m = 500$ experiments. Tables 4.5.10 – 4.5.12 record the SBGD_{pq} errors, measured by $\mathbb{E}|\mathbf{X}_{\text{SOL}} - \mathbf{x}_*|^2$, for different values of descent parameter, λ , and for different protocols of (p, q) . The sensitive dependence on λ is observed for both the non-communicating GD(BT) (see Table 4.5.7) and the SBGD_{pq} . Once again, the parameters $(p, q) = (2, 1/2)$ yield the most stable performance of SBGD_{pq} , whereas other scaling of GD(BT) and SBGD are more sensitive to the choice of λ .

In summary, Tables 4.5.8 – 4.5.12 indicate that while the results of SBGD_{pq} are mostly comparable, $\text{SBGD}_{(2,1/2)}$ seems to provide optimal results, with main emphasize on $p = 2$. At the same time, we conclude that the tuning parameters, (p, q) , has a limited effect on the overall performance of SBGD_{pq} method. Accordingly, we did not optimize these parameters.

Motivated by these findings, we focus below on two versions of SBGD (see also Tables 4.1.1 and 4.1.2): the basic version, $\text{SBGD}_{1,1}$ with $\phi_p(\eta) = \eta$ and $\phi_q(\tilde{m}) = \tilde{m}$, and $\text{SBGD}_{2,1}$ with

$\phi_p(\eta) = \eta^2$ which seems to be a universally better.

4.5.2.3 2D comparison with the Adam method

We report on the results of SBGD_{pq} , compared with the Adam method for the 2D Rastrigin function (4.5.2). We also include results for the $\text{GD}(h)$ and $\text{GD}(\text{BT})$ methods. As in the case of the 1D objective function (4.1.4), we distinguish between two cases: initial data uniformly distributed in $[-3, 3]^2$ enclosing the global minimum at the origin (Table 4.5.13), vs. initial data uniformly distributed at $[-3, -1]^2$ (Table 4.5.14). The SBGD_{pq} variants with $(p, q) = (1, 1)$ and $(p, q) = (2, 1)$ were computed for $m = 1000$ simulations with random with descent parameter $\lambda = 0.8$, and using tolerance parameters, $\text{tolmerge} = 0.1$, $\text{tolm} = 0.01$ and $\text{tolres} = 10^{-4}$.

The SBGD_{pq} variant with $(p, q) = (2, 1)$ seems to be consistently better than the primary version $(p, q) = (1, 1)$. When the methods are initiated at $[-3, 3]^2$, the SBGD variants provide comparable or better results than $\text{GD}(h)$, $\text{GD}(\text{BT})$ and Adams methods. When initiated at $[-3, -1]^2$ which does not enclose $\mathbf{x}_*$, the SBGD variants provide distinctively better results than $\text{GD}(h)$, $\text{GD}(\text{BT})$ and Adam methods. The Adam iterations with the smaller step size $\alpha = 0.2$ remain trapped in local basins of attraction.

4.5.3 20-dimensional problems

We now turn attention to the computation of the global minimizer for the 20-dimensional Rastrigin and Ackley functions. As the dimension of the ambient space to be explored increases, so does the number of agents, N , necessary to explore that space in order to secure a ‘faithful’ approximate minimizer. The increase of N dependence on the dimension d is intimately related to the way one quantifies the quality of such an approximation .

4.5.3.1 Success rate — N vs. d

One approach to measure success that was used in one- and two-dimensional problems, is to secure a computed solution within a pre-determined neighborhood of the global minimizer, $[\mathbf{x}_* - 0.25, \mathbf{x}_* + 0.25]^d$. This approach places severe restrictions in the case of high-dimensional data. Specifically, Table 4.5.15 and even more so, Table 4.5.16, show the rapid growth in the number of SBGD agents, $N = N(d)$, which are required to ensure 80% success rate in $m = 500$ tests of high-dimensional Ackley, and respectively, 70% success rate in Rastrigin benchmark functions. In both cases, one observes a rather small *critical dimension*, d_c , such that $N(d) \gg N(d_c)$ for $d > d_c$.

An alternative approach to quantify the quality of computed minimizers is to measure the expected (average) error in position, $\mathbb{E}[\|\mathbf{X}_{\text{SOL}} - \mathbf{x}_*\|]$. The results recorded in Tables 4.5.17 and 4.5.18 for 20-dimensional Ackley and, respectively, Rastrigin functions, indicate the advantage of SBGD₂₁ as the shift B increases. But the main point to observe is that even for relatively large $N = 100, 200$, the results fail to faithfully capture the global minimizer. Indeed, we claim that going beyond the two examples of Ackley and Rastrigin, measuring the (average) distance to of the computed minimizers, is not necessarily an effective quantifier for the quality of an optimizer: one might need a large $N = N(d)$ before approaching a small neighborhood of the global minimizer.

4.5.3.2 Measuring the loss

A more ‘faithful’ way to measure the quality of numerical optimizers in high-dimensional data is to measure the average loss (height), $\mathbb{E}[f(\mathbf{X}_{\text{SOL}})]$. After all, the underlying goal is to minimize the value of f . In particular, one might argue that $\mathbf{X}_{\text{SOL}}$ will provide a faithful approximation whenever $f(\mathbf{X}_{\text{SOL}}) - f(\mathbf{x}_*)$ is small, even if $\mathbf{X}_{\text{SOL}}$ remains far from $\mathbf{X}_i^n$. This approach of ‘looking from above’, masks the increasing complexity with the increasing dimension.

We begin, in Table 4.5.19, where we record the average ‘loss’ obtained by SBGD₂₁, GD(BT)

and Adam(0.5) with initial data equi-distributed in $[-3, 3]^2$ for the 20-dimensional Rastrigin function. SBGD₂₁ was used with descent parameter $\lambda = 0.8$, shrinkage parameter $\gamma = 0.5$ and thresholds

$$tolmerge = 10^{-1}, \quad tol m = 10^{-3}, \quad tolres = 10^{-2}.$$

The reason for the smaller shrinkage parameter $\gamma = 0.5$ was efficiency: backtracking is accelerated with more rough backtracking steps, yet this does not seem to deteriorate the quality of SBGD results. As indicated before, the results for SBGD₁₁ were only slightly worse than but otherwise comparable to SBGD₂₁ and therefore are not recorded here. We make two observations.

(i) For a small number of agents, $N \leq 10$, the results, particularly SBGD₂₁ and GD(BT) are comparable. Indeed, we recall that SBGD eliminates the lightest agents, so that after $N - 1$ iterations, it is left with a single agent which explores the large, uncharted ambient space, much like a single-agent method.

(ii) There is a clear trend that we saw before: SBGD₂₁ outperforms the non-communicating GD(BT) and Adam when the global minimum is *not* enclosed within the initial domain of initial data. While the results are comparable for $B = 0, 1$ and/or small N 's, there is an increasing difference for $B = 3, 5$ and $N > 20$.

Finally, in Table 4.5.20 we report on the corresponding comparison of average loss for the 20-dimensional Ackley function. Here, one encounters a much more sensitive dependence on the initial step size: we had to increase $h_0 = 2$ (instead of $h_0 = 1$ used before) in order to realize the advantage of SBGD₂₁. We maintain the usual backtracking parameters $\lambda = 0.2$ and $\gamma = 0.9$, and threshold parameters

$$tolmerge = 10^{-3}, \quad tol m = 10^{-4}, \quad tolres = 10^{-4}.$$

It is observed that for moderately shifted cases $B \leq 3$, SBGD₂₁ has clearly better performance for $N > 20$. However, as the shifting increases to $B = 5$, the benefit of communications is reduced, SBGD₂₁ becomes comparable with Adam(0.5).

4.5.3.3 SBGD as pre-conditioner

A more practical strategy is to take the expected value, $\mathbb{E}[\mathbf{x}_{SOL}]$, varying overall SBGD solutions resulted from m randomly generated initial configurations, as a good initial guess and then iterate it with the steepest descent correction. In this way, the algorithm is expected to mimic the convergence in expectation $\mathbb{E}[\mathbf{X}_{SOL}] \xrightarrow{n \rightarrow \infty} \mathbf{x}_*$. Table 4.5.21 records the L^∞ –distance between the expectation $\mathbb{E}[\mathbf{X}_{SOL}]$ and the global minimizer $\mathbf{x}_*$ for the 20-dimensional Ackley and Rastrigin functions. The expectation is computed with $m = 1000$ runs. We also present the error of the corrected solution $\mathbf{x}_{corr}$, which is obtained by iterating $\mathbb{E}[\mathbf{X}_{SOL}]$ with gradient descent until $|\nabla f(\mathbf{x}_{corr})|_2 < 10^{-3}$. For both benchmark functions, this strategy gives very good solutions with a reasonable amount of agents. The quality of $\mathbb{E}[\mathbf{X}_{SOL}]$ is improved with more agents applied.

We also investigate the effect of shifting in the initial distribution. The solution $\mathbf{x}_{corr}$ is considered to be a correct approximation of global minimizer if it lies in the region $Q_{0.5}(\mathbf{x}_*)$. Table 4.5.22 and 4.5.23 show the results for the Ackley and the Rastrigin function under varying shift parameters and different agent numbers. It turns out that the quality of the solution is very sensitive to the initial data in high dimension. The algorithm fails to work correctly in the Ackley test cases when $B = 2$. The situation is even worse for the Rastrigin test cases due to the unclear difference between different minimums. The method gives wrong solutions in all the test cases when B is 1.5.

The results above indicate that the solutions obtained for high-dimensional problems can be significantly affected by the shifting in the target function, especially when the global minimum is not very distinguishable from the other minima.

$\phi_p(\eta)\psi_q(\tilde{m})$	$\sqrt{\tilde{m}}$	$\tilde{m}$	$\tilde{m}^2$
η	2.53e-01	2.07e-01	1.78e-01
η^2	8.51e-02	3.07e-01	1.70e-01
η^{20}	6.39e-02	2.01e-01	2.74e-01
η^{50}	5.75e-02	1.88e-01	2.61e-01

Table 4.5.10: 2D Rosenbrock, SBGD_{pq}
with $\lambda = 0.3$.

$\phi_p(\eta)\psi_q(\tilde{m})$	$\sqrt{\tilde{m}}$	$\tilde{m}$	$\tilde{m}^2$
η	6.79e-02	6.18e-02	6.05e-02
η^2	4.97e-02	7.10e-02	6.08e-02
η^{20}	4.65e-02	5.95e-02	6.74e-02
η^{50}	4.98e-02	6.02e-02	6.87e-02

Table 4.5.11: 2D Rosenbrock, SBGD_{pq}
with $\lambda = 0.5$.

$\phi_p(\eta)\psi_q(\tilde{m})$	$\sqrt{\tilde{m}}$	$\tilde{m}$	$\tilde{m}^2$
η	7.11e-02	1.51e-01	1.29e-01
η^2	6.83e-02	1.22e-01	1.29e-01
η^{20}	1.21e-01	6.45e-02	1.48e-01
η^{50}	1.26e-01	6.76e-02	1.37e-01

Table 4.5.12: 2D Rosenbrock, SBGD_{pq}
with $\lambda = 0.7$.

N	5	10	15	20	30
SBGD ₁₁	34.4%	52.1%	62.6%	70.0%	75.8%
SBGD ₂₁	34.5%	60.1%	75.3%	84.3%	91.0%
GD(0.004)	36.3%	50.5%	60.0%	70.0%	78.1%
GD(BT)	35.0%	51.0%	62.0%	70.8%	79.3%
Adam(0.8)	23.7%	29.6%	39.1%	46.8%	65.5%
Adam(0.2)	32.1%	40.9%	55.9%	65.3%	79.4%

Table 4.5.13: Success rates of SBGD compared with GD(h) , GD(BT) and Adam methods for global optimization of (4.5.2)($B = 0$), based on $m = 1000$ runs with uniformly generated initial data in $[-3, 3]^2$.

N	5	10	15	20	30
SBGD ₁₁	17.0%	49.2%	61.7%	67.0%	72.7%
SBGD ₂₁	14.2%	46.7%	68.4%	81.9%	89.6%
GD(0.004)	0.0%	0.0%	0.0%	0.0%	0.0%
GD(BT)	1.8%	2.4%	3.4%	4.3%	5.9%
Adam(0.8)	24.5%	31.3%	41.4%	49.2%	66.9%
Adam(0.2)	0.0%	0.0%	0.0%	0.0%	0.0%

Table 4.5.14: Success rates of SBGD compared with GD(h) , GD(BT) and Adam methods for global optimization of 2D Rastrigin (4.5.2)($B = 0$) based on $m = 1000$ runs of uniformly generated initial data in $[-3, -1]^2$.

d	10	11	12	13	14	15	16
N	15	18	23	42	120	540	3000

Table 4.5.15: 80% success rate with 20D Ackley function

d	1	2	3	4
N	4	23	180	2900

Table 4.5.16: 70% success rate with 20D Rastrigin function

$\mathbf{x}_* = B$		N=50	N=100	N=200
$B = 0$	SBGD ₂₁	9.00e-07	2.02e-07	1.43e-07
	GD(BT)	1.18e-01	6.90e-02	1.88e-02
	Adam(0.5)	3.37e-03	4.03e-03	4.96e-03
$B = 3$	SBGD ₂₁	1.65e-06	1.51e-06	7.96e-07
	GD(BT)	7.64	6.72	5.67
	Adam(0.5)	2.14e-01	1.22e-01	1.27e-01
$B = 5$	SBGD ₂₁	4.15	1.07	2.36e-01
	GD(BT)	17.99	17.59	17.11
	Adam(0.5)	11.01	9.27	7.52

Table 4.5.17: $\mathbb{E}[\mathbf{X}_{SOL} - \mathbf{x}_*]$ for 20-dimensional shifted Ackley. $m = 1000$.

$\mathbf{x}_* = B$		N=50	N=100	N=200
$B = 0$	SBGD ₂₁	4.03e-01	4.98e-01	3.91e-01
	GD(BT)	2.96e-01	3.25e-01	4.35e-01
	Adam(0.5)	3.75e-01	2.86e-01	4.14e-01
$B = 3$	SBGD ₂₁	5.07	3.71	2.92
	GD(BT)	9.43	9.02	8.57
	Adam(0.5)	9.02	8.63	8.15
$B = 5$	SBGD ₂₁	3.53	2.45	1.29
	GD(BT)	18.02	17.64	17.18
	Adam(0.5)	17.13	16.73	16.3

Table 4.5.18: $\mathbb{E}[\mathbf{X}_{SOL} - \mathbf{x}_*]$ for 20-dimensional shifted Rastrigin. $m = 1000$.

$\mathbf{x}_* = B$		N=5	N=10	N=20	N=50	N=100	N=200
$B = 0$	SBGD ₂₁	54.30	43.74	39.34	34.55	32.01	33.95
	GD(BT)	53.43	44.62	39.43	34.55	31.85	29.53
	Adam(0.5)	48.66	44.02	37.35	32.65	30.46	27.76
$B = 1$	SBGD ₂₁	65.75	53.24	47.14	40.90	37.20	33.95
	GD(BT)	64.86	53.98	47.1	40.77	37.06	33.57
	Adam(0.5)	56.02	50.35	42.56	37.69	34.58	31.79
$B = 3$	SBGD ₂₁	189.24	166.46	149.14	121.45	102.98	94.77
	GD(BT)	187.8	167.58	152.46	122.35	125.44	115.83
	Adam(0.5)	164.46	146.82	137.16	122.84	115.16	105.5
$B = 5$	SBGD ₂₁	463.09	387.88	254.23	200.34	160.62	142.22
	GD(BT)	465.42	434.9	409.11	381.4	362.54	345.52
	Adam(0.5)	411.77	381.19	364.68	339.81	325.86	309.99

Table 4.5.19: 20D Rastrigin. Average loss with $m = 1000$ simulations.

$\mathbf{x}_* = B$		N=10	N=20	N=50	N=100	N=200
$B = 0$	SBGD ₂₁	0.89	2.62e-05	1.17e-05	9.72e-07	1.82e-07
	GD(BT)	2.76	1.92	0.97	0.41	7.73e-02
	Adam(0.5)	0.37	0.073	0.061	0.061	0.061
$B = 1$	SBGD ₂₁	2.38	0.27	3.9e-05	6e-06	1e-06
	GD(BT)	3.01	2.00	0.92	0.31	5.45e-02
	Adam(0.5)	0.87	0.17	0.061	0.061	0.061
$B = 3$	SBGD ₂₁	7.67	4.41	0.72	0.06	5.4e-05
	GD(BT)	7.91	7.47	6.81	6.13	5.37
	Adam(0.5)	3.74	2.65	1.70	1.22	0.79
$B = 5$	SBGD ₂₁	12.01	11.49	10.01	8.47	7.09
	GD(BT)	11.98	11.8	11.53	11.36	11.17
	Adam(0.5)	10.72	9.70	8.45	7.58	6.67

Table 4.5.20: 20D Ackley. Average loss with $m = 1000$ simulations.

		N=50	N=100	N=200
Ackley function	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	$8.29e^{-2}$	$7.28e^{-2}$	$6.79e^{-2}$
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	$2.07e^{-13}$	$1.96e^{-13}$	$2.18e^{-13}$
Rastrigin function	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	$2.07e^{-1}$	$7.86e^{-2}$	$7.36e^{-2}$
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	$9.06e^{-6}$	$2.29e^{-6}$	$2.05e^{-6}$

Table 4.5.21: Errors of the expectation and the corrected solution in $d = 20$, $B = C = 0$.

$\mathbf{x}_* = B$		N=50	N=100	N=200
$B = 1$	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	$3.97e^{-1}$	$3.53e^{-1}$	$3.36e^{-2}$
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	$2.10e^{-13}$	$1.29e^{-13}$	$1.15e^{-13}$
$B = 1.5$	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	$7.45e^{-1}$	$6.78e^{-1}$	$5.43e^{-1}$
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	$9.69e^{-1}$	$1.27e^{-13}$	$1.60e^{-13}$
$B = 2$	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	1.20	1.09	1.01
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	$9.69e^{-1}$	$9.69e^{-1}$	$9.68e^{-1}$

Table 4.5.22: Solutions for the shifted Ackley function by SGBD, $d = 20$, $m = 1000$.

$\mathbf{x}_* = B$		N=50	N=100	N=200
$B = 0.5$	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	$2.21e^{-1}$	$2.38e^{-1}$	$2.24e^{-1}$
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	$5.86e^{-6}$	$1.06e^{-5}$	$6.53e^{-6}$
$B = 1$	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	$5.97e^{-1}$	$4.80e^{-1}$	$4.19e^{-1}$
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	1.99	$5.89e^{-6}$	$1.11e^{-5}$
$B = 1.5$	$ \mathbb{E}[\mathbf{X}_{SOL}] - \mathbf{x}_* _\infty$	$7.92e^{-1}$	$7.00e^{-1}$	$6.34e^{-1}$
	$ \mathbf{x}_{corr} - \mathbf{x}_* _\infty$	$9.95e^{-1}$	$9.95e^{-1}$	1.99

Table 4.5.23: Solutions for the shifted Rastrigin function by SGBD, $d = 20$, $m = 1000$.

Chapter 5: Conclusions

5.1 Summary

In this thesis, we have discussed the swarm dynamics arising from flocking phenomena and the application to numerical optimization. At first, we studied the large time behavior of the celebrated Cucker-Smale system. In this agent-based model, motions of particles are dictated by the alignment of velocities towards the average heading of neighbors. Cross-individual interactions are characterized by a uniform or an array of distinct communication kernels, dividing the scenario into single-species and, respectively, multi-species. With long-range communications satisfying fat-tail connectivity, an important outcome is the flocking dynamics, namely, the spatial distribution of the crowd remains bounded, and, the velocities of agents are asymptotically aligned with a constant-in-time bulk mean velocity. It is of interest to investigate the preservation of flocking behavior as we take hydrodynamic limit, where one encounters the lack of closure for pressure. In contrast to the previous studies where vanishing pressure is usually assumed, the main contribution of our work is to provide the very first proof on the hydrodynamic alignment for multi-species associated with *general* entropic pressure. The proof employs the network connectivity and a weak dispersion bound. As a key element of recovering specific flocking rate, the boundedness of velocities was addressed for one-dimensional multi-species and the extended p -alignment dynamics.

To numerically simulate collective dynamics, general options include particle (agent-based) simulations, where the motion of each agent is traced, and hydrodynamics simulations, which provides a more economical approach when a large population is involved. Due to the fact that

the emergence of collective behavior is mostly dependent on the communication protocol instead of the accuracy of numerical scheme, efficient and stability-preserving numerical methods would be preferred. For particle simulations, we proved that the simple forward Euler time stepping ensures the decay of energy fluctuations as applied to the Cucker-Smale dynamics. And it recovered the expected flocking rate in the simulations of single-species. Hydrodynamics simulations involves solving nonlinear transport equations. The nonlinearity may lead to the development of discontinuity even if initial data is smooth. To ensure non-oscillatory numerical solutions with a good accuracy, we suggested using the second order central schemes. These schemes can generate high-resolution solutions for conservation laws with high efficiency by avoiding local characteristic decompositions. A key ingredient is to employ piecewise linear MUSCL reconstruction with proper slope limiter. While it is long believed that the minmod structure is necessary for TVD limiters, we proved that the smooth van Albada slope limiter, despite the lack of minmod structure, also ensures theoretical TVD stability as applied to the fully discrete NT schemes and the semi-discrete MUSCL schemes. In numerical experiments, the improved resolution of the van Albada limiter over the classical minmod limiter is shown through the test cases for conservation laws and incompressible Euler equations. Computations of one- and two-dimensional flocking dynamics have verified the validity of hydrodynamic approach.

Besides the flocking-based swarm dynamics, we have also studied the swarm-based approach for non-convex optimization. Classical gradient-based optimization methods often suffer from getting trapped at local minima due to the limited local information. To enhance the perception of global landscape, the strategy of swarm-based algorithms is to employ a population of interacting agents to search the solution collectively. The innovation of our work is to propose a novel swarm-based gradient descent (SBGD) method. The dynamics employs a group of agents, each of which conducts local gradient descent search and exchanges information with each other. The new SBGD is distinguished from the other existing swarm-based algorithms by its novel mass-based communications: each agent is identified by not only its position, $\mathbf{x}$, but also an individual mass, m , which indicates the quality of agent. At each iteration, the best minimizer

gains weights from the other agents, gradually becoming the ‘heavy’ leader of the swarm. This allows dynamical, self-adjusting control on individual search modes: the good (heavy) agents will take small time-step to search the surrounding carefully, as they are suspected to be close to the (near-)optimal minimizer, whereas the bad (light) agents will take more aggressive step size to explore the region far way from the current positions, during which they may encounter better minimizing areas. With the properly constructed time-step protocol based on backtracking line search, SBGD enlarges the range of exploration of original gradient descent methods, while the rigorously proved convergence is still maintained. Numerical comparison with the non-communicating GD and Adam methods was conducted in several one-, two-, and 20-dimensional benchmarks, the results have shown the great competitiveness of SBGD in non-convex problems.

5.2 Future works

We list below some potential aspects of future works.

Short-range communication. In the existing discussions on Cucker-Smale dynamics, it is usually assumed that the communication kernel ϕ admits a ‘fat-tail’ and therefore is non-vanishing on the whole space. However, in real world modeling the communications may only occur between neighboring individuals, i.e. ϕ is compactly supported. Analytical study on short-range communications requires further investigations.

Affordable moment closure. Although the hydrodynamic alignment driven by Cucker-Smale system was analytically proved without assuming specific thermodynamic closure, for concrete simulations it is still necessary to close the highest moment. For general social dynamics, computing the exact mesoscopic distribution involves solving high-dimensional Vlasov-Fokker-Planck equation(s), which exceeds the current computational capability. Hence, affordable approximate methods would be required for moment closure. An appealing direction is data-driven approach. For instance, in the series of works by Juntao Huang et al [65, 66, 67], a machine learning tech-

nique was proposed for the radiative transfer equation, which learns the gradient of the highest moment as a function of lower-order moments. Similar idea may be extended to general kinetic descriptions. Alternatively, dynamical low rank approximation e.g. [27, 38, 81] may also provide a computationally tractable approach to recover kinetic information.

Swarm-based optimizers with masses. The uniqueness of SBGD algorithm lies in the idea of introducing individual masses to distinguish different agents. In view of the remarkable improvement of SBGD against the original gradient descent methods, an appealing future work would be to investigate the validity of employing mass-based communications to improve other gradient-based algorithms, e.g. conjugate gradient and Adam method. Similar approach may also be applied to improve gradient-free swarm-based methods e.g. PSO and CBO. Indeed, gradient-free methods can avoid expensive gradient evaluations in complicated problems and may be more resilient to complex geometry in high dimensions. By further allowing adaptive control on individual search parameters using masses, the algorithm may better ensure the stability of convergence under more relaxed initial parameter constraints.

Appendix A: Proof of Theorem 2.1.5

Proof of Theorem 2.1.5. We argue along the lines of [57]. Consider the equivalent representation of momentum equations,

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} = \int \phi(|\mathbf{x} - \mathbf{x}'|)(\mathbf{u}(t, \mathbf{x}') - \mathbf{u}(t, \mathbf{x}))\rho(t, \mathbf{x}') d\mathbf{x}'. \quad (\text{A.0.1})$$

We fix an arbitrary unit vector $\mathbf{w} \in \mathbb{R}^d$. Projection of (A.0.1) onto the space spanned by $\mathbf{w}$ yields

$$(\partial_t + \mathbf{u} \cdot \nabla) \langle \mathbf{u}(t, \mathbf{x}), \mathbf{w} \rangle = \int \phi(|\mathbf{x} - \mathbf{x}'|)(\langle \mathbf{u}(t, \mathbf{x}'), \mathbf{w} \rangle - \langle \mathbf{u}(t, \mathbf{x}), \mathbf{w} \rangle)\rho(t, \mathbf{x}') d\mathbf{x}' \quad (\text{A.0.2})$$

Assume that $\langle \mathbf{u}(t, \mathbf{x}), \mathbf{w} \rangle$ reaches a maximum at $(\mathbf{x}_+, t)$ and a minimum at $(\mathbf{x}_-, t)$. Denote

$$u_+(t) := \max_{\mathbf{x} \in \mathcal{S}(t)} \langle \mathbf{u}(t, \mathbf{x}), \mathbf{w} \rangle = \langle \mathbf{u}(t, \mathbf{x}_+), \mathbf{w} \rangle.$$

It follows that $u_+(t)$ satisfies

$$\begin{aligned} \frac{d}{dt} u_+(t) &= \int \phi(|\mathbf{x}_+ - \mathbf{x}'|)(\langle \mathbf{u}(t, \mathbf{x}'), \mathbf{w} \rangle - \langle \mathbf{u}(t, \mathbf{x}_+), \mathbf{w} \rangle)\rho(t, \mathbf{x}') d\mathbf{x}' \\ &\leq \phi(D(t)) \int (\langle \mathbf{u}(t, \mathbf{x}'), \mathbf{w} \rangle - \langle \mathbf{u}(t, \mathbf{x}_+), \mathbf{w} \rangle)\rho(t, \mathbf{x}') d\mathbf{x}'. \end{aligned}$$

Similarly, $u_-(t) := \min_{\mathbf{x} \in \mathcal{S}(t)} \langle \mathbf{u}(t, \mathbf{x}), \mathbf{w} \rangle$ admits the lower bound

$$\frac{d}{dt} u_-(t) \geq \phi(D(t)) \int (\langle \mathbf{u}(t, \mathbf{x}'), \mathbf{w} \rangle - \langle \mathbf{u}(t, \mathbf{x}_-), \mathbf{w} \rangle)\rho(t, \mathbf{x}') d\mathbf{x}'$$

The difference of the two inequalities implies

$$\frac{d}{dt}|u_+(t) - u_-(t)| \leq -\phi(D(t))M|u_+(t) - u_-(t)|.$$

Since the above inequality holds for an arbitrary $|\mathbf{w}| = 1$, it follows that the velocity fluctuation,

$\delta V(t) = \sup_{|\mathbf{w}|=1} |u_+(t) - u_-(t)|$, satisfies

$$\frac{d}{dt}\delta V(t) \leq -\phi(D(t))M \cdot \delta V(t) \lesssim -M(1 + D(t))^{-\theta} \cdot \delta V(t). \quad (\text{A.0.3})$$

An immediate consequence is that the pointwise velocity fluctuation does not increase

$$\frac{d}{dt}\delta V(t) \leq 0 \implies \delta V(t) \leq \delta V_0.$$

Subsequently, the spatial diameter grows no faster than linearly,

$$\frac{d}{dt}D(t) \leq \delta V(t) \implies D(t) \leq D_0 + \int_0^t \delta V(s)ds \leq D_0 + t\delta V_0. \quad (\text{A.0.4})$$

We prove the statements for $\theta = 1$ and $\theta < 1$ separately.

Case 1: $\theta = 1$. It remains to prove that $\delta V(t)$ decays with a polynomial rate. Substituting (A.0.4) into (A.0.3) yields

$$\frac{d}{dt}\delta V(t) \lesssim -M(1 + D_0 + \delta V_0 t)^{-1} \cdot \delta V(t)$$

We use Grownwall's inequality to obtain the desired estimate,

$$\delta V(t) \lesssim \delta V_0 \cdot (1 + t)^{-\nu_1}, \quad \nu_1 = M(\max\{1 + D_0, \delta V_0\})^{-1}. \quad (\text{A.0.5})$$

Case 2: $\theta < 1$. We prove the exponential decay of $\delta V(t)$ and the uniform boundedness of $D(t)$ through a bootstrapping argument.

Step 1. Fractional exponential decay. We first obtain a weak fractional exponential decay for $\delta V(t)$. Using (A.0.3) and (A.0.4) we find

$$\frac{d}{dt}\delta V(t) \lesssim -M(1 + D_0 + \delta V_0 t)^{-\theta} \cdot \delta V(t).$$

Grownwall's inequality yields

$$\delta V(t) \lesssim \delta V_0 \cdot e^{-\nu' t^{1-\theta}}, \quad \nu' = \frac{M}{(1-\theta)\delta V_0^\theta}. \quad (\text{A.0.6})$$

Step 2. Finite diameter. The fractional exponential decay (A.0.6) implies a bounded spatial diameter in time,

$$\frac{d}{dt}D(t) \leq \delta V(t) \lesssim \delta V_0 \cdot e^{-\nu' t^{1-\theta}} \implies D(t) \leq D_\infty := D_0 + C \cdot \delta V_0 < \infty. \quad (\text{A.0.7})$$

Here the constant, $C := \int_0^\infty e^{-\nu' t^{1-\theta}} dt < \infty$, depends on $M, \theta, \delta V_0$.

Step 3. Exponential decay. The uniform bound of $D(t)$ implies a uniform lower bound on the minimal communication, $\phi(D(t)) \geq \phi(D_\infty)$. Using (A.0.3) we obtain the exponential decay,

$$\delta V(t) \leq \delta V(0)e^{-\nu_2 t}, \quad \nu_2 = M\phi(D_\infty). \quad (\text{A.0.8})$$

□

Appendix B: Derivation of Multi-Species Hydrodynamics

We derive the mesoscopic and hydrodynamic descriptions of multi-species following the lines of [58].

From agent-based to hydrodynamic description

We formulate the time evolution of each probability density f_α , as defined in (2.2.5), based on the microscopic agent-based model (2.2.4). To this end, we test $\partial_t f_\alpha$ against arbitrary smooth function φ

$$\begin{aligned}
 \iint \partial_t f_\alpha(\mathbf{x}, \mathbf{v}, t) \varphi(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} &= \frac{d}{dt} \left[\frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} \varphi(\mathbf{x}_\alpha^i(t), \mathbf{v}_\alpha^i(t)) \right] \\
 &= \frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} [\dot{\mathbf{x}}_\alpha^i \cdot \nabla_{\mathbf{x}} \varphi(\mathbf{x}_\alpha^i, \mathbf{v}_\alpha^i) + \dot{\mathbf{v}}_\alpha^i \cdot \nabla_{\mathbf{v}} \varphi(\mathbf{x}_\alpha^i, \mathbf{v}_\alpha^i)] \quad (\text{B.0.1}) \\
 &= \frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} [\mathbf{v}_\alpha^i \cdot \nabla_{\mathbf{x}} \varphi(\mathbf{x}_\alpha^i, \mathbf{v}_\alpha^i) + F_\alpha^i \cdot \nabla_{\mathbf{v}} \varphi(\mathbf{x}_\alpha^i, \mathbf{v}_\alpha^i)].
 \end{aligned}$$

Here the alignment forcing F_α^i is given by

$$F_\alpha^i = \frac{d}{dt} \mathbf{v}_\alpha^i = \sum_{\beta \in I} \frac{1}{N_\beta} \sum_{j=1}^{N_\beta} \phi_{\alpha\beta}(|\mathbf{x}_\beta^j - \mathbf{x}_\alpha^i|) (\mathbf{v}_\beta^j - \mathbf{v}_\alpha^i) = \sum_{\beta \in I} L_{\alpha\beta}[f_\beta](\mathbf{x}_\alpha^i, \mathbf{v}_\alpha^i),$$

where $L_{\alpha\beta}[f_\beta](\mathbf{x}_\alpha^i, \mathbf{v}_\alpha^i) = \int \phi_{\alpha\beta}(|\mathbf{x}' - \mathbf{x}_\alpha^i|)(\mathbf{w} - \mathbf{v}_\alpha^i)f_\beta(\mathbf{x}', \mathbf{w})d\mathbf{x}'d\mathbf{w}$. Formal integration by part of (B.0.1) yields

$$\begin{aligned} \iint \partial_t f_\alpha(\mathbf{x}, \mathbf{v}, t) \varphi(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} &= \iint [\mathbf{v} \cdot \nabla_{\mathbf{x}} \varphi(\mathbf{x}, \mathbf{v}) + \sum_{\beta \in I} L_{\alpha\beta}[f_\beta] \cdot \nabla_{\mathbf{v}} \varphi(\mathbf{x}, \mathbf{v})] f_\alpha(\mathbf{x}, \mathbf{v}) d\mathbf{x} d\mathbf{v} \\ &= - \iint [\mathbf{v} \cdot \nabla_{\mathbf{x}} f_\alpha + \nabla_{\mathbf{v}} \cdot (\sum_{\beta \in I} L_{\alpha\beta}[f_\beta] f_\alpha)] \varphi d\mathbf{x} d\mathbf{v}. \end{aligned}$$

Since the above integral equation holds for arbitrary test function φ , we obtain the *mesoscopic* description of multi-species

$$\partial_t f_\alpha + \mathbf{v} \cdot \nabla_{\mathbf{x}} f_\alpha + \nabla_{\mathbf{v}} \cdot (\sum_{\beta \in I} L_{\alpha\beta}[f_\beta] f_\alpha) = 0. \quad (\text{B.0.2})$$

The bi-linear expression $\sum_{\beta \in I} L_{\alpha\beta}[f_\beta] f_\alpha$ represents the alignment interactions between different species.

Integration of the mesoscopic equation (B.0.2) in velocity variable $\mathbf{v}$ yields the mass equation

$$\partial_t \rho_\alpha + \nabla_{\mathbf{x}} \cdot (\rho_\alpha \mathbf{u}_\alpha) = 0, \quad \forall \alpha \in \mathcal{I}.$$

The momentum equations are obtained by integrating (B.0.2) against $\mathbf{v}$

$$0 = \int [\partial_t(\mathbf{v} f_\alpha) + \mathbf{v}(\mathbf{v} \cdot \nabla_{\mathbf{x}} f_\alpha) + \mathbf{v} \nabla_{\mathbf{v}} \cdot (\sum_{\beta \in \mathcal{I}} L_{\alpha\beta}(f_\beta) f_\alpha)] d\mathbf{v} = I + II + III. \quad (\text{B.0.3})$$

The first term is the time derivative of momentum density, $\rho_\alpha \mathbf{u}_\alpha$, in (2.2.6)

$$I = \partial_t(\rho_\alpha \mathbf{u}_\alpha). \quad (\text{B.0.4})$$

The second term can be written as

$$\begin{aligned}
II &= \nabla_{\mathbf{x}} \cdot (\rho_{\alpha} \mathbf{u}_{\alpha} \otimes \mathbf{u}_{\alpha}) + \nabla_{\mathbf{x}} \cdot \int (\mathbf{v} - \mathbf{u}_{\alpha}) \otimes (\mathbf{v} - \mathbf{u}_{\alpha}) f_{\alpha}(\mathbf{x}, \mathbf{v}) d\mathbf{v} \\
&:= \nabla_{\mathbf{x}} \cdot (\rho_{\alpha} \mathbf{u}_{\alpha} \otimes \mathbf{u}_{\alpha}) + \nabla_{\mathbf{x}} \cdot \mathbb{P}_{\alpha},
\end{aligned} \tag{B.0.5}$$

where the pressure tensor $\mathbb{P}_{\alpha}$ is as defined in (2.2.7). For the third term, integration by parts yields

$$\begin{aligned}
III &= \int \mathbf{v} \nabla_{\mathbf{v}} \cdot \left(\sum_{\beta \in \mathcal{I}} L_{\alpha\beta}[f_{\beta}] f_{\alpha} \right) d\mathbf{v} = - \sum_{\beta \in \mathcal{I}} \int L_{\alpha\beta}[f_{\beta}] f_{\alpha} d\mathbf{v} \\
&= - \sum_{\beta \in \mathcal{I}} \iint \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (\mathbf{u}_{\beta}(\mathbf{x}') - \mathbf{u}_{\alpha}(\mathbf{x})) \rho_{\alpha}(\mathbf{x}) \rho_{\beta}(\mathbf{x}') d\mathbf{x}' = -\mathbf{A}_{\alpha},
\end{aligned} \tag{B.0.6}$$

where $\mathbf{A}_{\alpha}$ is the inter-species alignment in (2.2.2). Substituting (B.0.4)–(B.0.6) into (B.0.3) results in the dynamics of momentum in (2.2.1).

Balance of energy

The balance of energy is obtained by integrating (B.0.2) against $\frac{|\mathbf{v}|^2}{2}$,

$$0 = \int \left[\partial_t \left(\frac{|\mathbf{v}|^2}{2} f_{\alpha} \right) + \frac{|\mathbf{v}|^2}{2} (\mathbf{v} \cdot \nabla_{\mathbf{x}} f_{\alpha}) + \frac{|\mathbf{v}|^2}{2} \nabla_{\mathbf{v}} \cdot \left(\sum_{\beta \in \mathcal{I}} L_{\alpha\beta}[f_{\beta}] f_{\alpha} \right) \right] d\mathbf{v} = I + II + III. \tag{B.0.7}$$

The first term is the time derivative of energy density $\rho_{\alpha} E_{\alpha}$

$$I = \partial_t (\rho_{\alpha} E_{\alpha}). \tag{B.0.8}$$

The second term can be written as

$$\begin{aligned}
II &= \nabla_{\mathbf{x}} \cdot \int \frac{|\mathbf{v}|^2}{2} \mathbf{v} f_{\alpha} d\mathbf{v} = \nabla_{\mathbf{x}} \cdot \int \frac{|\mathbf{v}|^2}{2} \mathbf{u}_{\alpha} f_{\alpha} d\mathbf{v} + \nabla_{\mathbf{x}} \cdot \int \frac{|\mathbf{v}|^2}{2} (\mathbf{v} - \mathbf{u}_{\alpha}) f_{\alpha} d\mathbf{v} \\
&= \nabla_{\mathbf{x}} \cdot (\rho_{\alpha} E_{\alpha} \mathbf{u}_{\alpha}) + \nabla_{\mathbf{x}} \cdot \int \frac{|\mathbf{v}|^2}{2} (\mathbf{v} - \mathbf{u}_{\alpha}) f_{\alpha} d\mathbf{v} \\
&= \nabla_{\mathbf{x}} \cdot (\rho_{\alpha} E_{\alpha} \mathbf{u}_{\alpha}) + \nabla_{\mathbf{x}} \cdot \int (\mathbf{v} - \mathbf{u}_{\alpha}) \otimes (\mathbf{v} - \mathbf{u}_{\alpha}) \mathbf{u}_{\alpha} f_{\alpha} d\mathbf{v} \\
&\quad + \nabla_{\mathbf{x}} \cdot \int \frac{|\mathbf{v} - \mathbf{u}_{\alpha}|^2}{2} (\mathbf{v} - \mathbf{u}_{\alpha}) f_{\alpha} d\mathbf{v} \\
&= \nabla_{\mathbf{x}} \cdot (\rho_{\alpha} E_{\alpha} + \mathbb{P}_{\alpha} \mathbf{u}_{\alpha} + \mathbf{q}_{\alpha}),
\end{aligned} \tag{B.0.9}$$

where the heat flux $\mathbf{q}_{\alpha} := \int \frac{|\mathbf{v} - \mathbf{u}_{\alpha}|^2}{2} (\mathbf{v} - \mathbf{u}_{\alpha}) f_{\alpha} d\mathbf{v}$. For the third term, integration by part yields

$$\begin{aligned}
III &= \int \frac{|\mathbf{v}|^2}{2} \nabla_{\mathbf{v}} \cdot \left(\sum_{\beta \in \mathcal{I}} L_{\alpha\beta} [f_{\beta}] f_{\alpha} \right) d\mathbf{v} = - \int \mathbf{v} \cdot L_{\alpha\beta} [f_{\beta}] f_{\alpha} d\mathbf{v} \\
&= - \sum_{\beta \in \mathcal{I}} \iiint \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') \mathbf{v} \cdot (\mathbf{v} - \mathbf{w}) f_{\alpha}(t, \mathbf{x}, \mathbf{v}) f_{\beta}(t, \mathbf{x}', \mathbf{w}) d\mathbf{x}' d\mathbf{w} d\mathbf{v} \\
&= - \sum_{\beta \in \mathcal{I}} \int \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (2E_{\alpha}(t, \mathbf{x}) - \mathbf{u}_{\alpha}(t, \mathbf{x}) \cdot \mathbf{u}_{\beta}(t, \mathbf{x}')) \rho_{\alpha}(t, \mathbf{x}) \rho_{\beta}(t, \mathbf{x}') d\mathbf{x}'.
\end{aligned} \tag{B.0.10}$$

Combining (B.0.8)–(B.0.10) we obtain the energy equation

$$\begin{aligned}
\partial_t(\rho_{\alpha} E_{\alpha}) + \nabla_{\mathbf{x}} \cdot (\rho_{\alpha} E_{\alpha} \mathbf{u}_{\alpha} + \mathbb{P}_{\alpha} \mathbf{u}_{\alpha} + \mathbf{q}_{\alpha}) = \\
- \sum_{\beta \in \mathcal{I}} \int \phi_{\alpha\beta}(\mathbf{x}, \mathbf{x}') (2E_{\alpha}(\mathbf{x}) - \mathbf{u}_{\alpha}(\mathbf{x}) \cdot \mathbf{u}_{\beta}(\mathbf{x}')) \rho_{\alpha}(\mathbf{x}) \rho_{\beta}(\mathbf{x}') d\mathbf{x}'.
\end{aligned}$$

Appendix C: Improved Dispersion Bound for One-Dimensional Multi-Species

Derivation of the improved dispersion bound (2.2.34) follows the lines of [95, Appendix C].

We start with the bound

$$D(t) \leq C_D(1+t)^\gamma, \quad \gamma = \gamma_s = \frac{1}{3-2s}.$$

Applying estimate (2.2.13) leads to flocking behavior with fractional exponential decay rate, which can be rewritten in the form

$$\delta \mathcal{E}(t) \leq \frac{1}{\chi(t)} \delta \mathcal{E}(0), \quad \chi(t) = \exp\{\nu_\theta t^{1-\gamma_\theta}\}.$$

Combine with (2.2.29), we have

$$\begin{aligned} \frac{d}{dt} \chi(t) \delta \mathcal{E}(t) &= \chi(t) \frac{d}{dt} \delta \mathcal{E}(t) + \dot{\chi}(t) \delta \mathcal{E}(t) \\ &\leq -M \rho_-^2 \chi(t) \sum_{\alpha} \|u_{\alpha}(t, \cdot)\|_{H^2(\mathcal{S}_{\alpha})}^2 + \frac{\dot{\chi}(t)}{\chi(t)} \delta \mathcal{E}(0) \\ &= -M \rho_-^2 \chi(t) \sum_{\alpha} \|u_{\alpha}(t, \cdot)\|_{H^2(\mathcal{S}_{\alpha})}^2 + \nu_\theta (1 - \gamma_\theta) t^{-\gamma_\theta} \delta \mathcal{E}(0). \end{aligned}$$

This implies

$$\int_0^t \chi(\tau) \sum_{\alpha} \|u_{\alpha}(\tau, \cdot)\|_{H^s(\mathcal{S}_{\alpha})}^2 d\tau \leq \frac{C_\rho^2}{M} \delta \mathcal{E}(0) + C_1 t^{1-\gamma_\theta} \delta \mathcal{E}(0), \quad C_1 = \frac{1}{M} C_\rho^2 \nu_\theta.$$

Revisit (2.2.32) with the last weighted bound, we obtain

$$\begin{aligned} D^{\frac{3}{2}-s}(t) &\leq \sum_a D_{\alpha 0}^{\frac{3}{2}-s} + \left(\int_0^t \chi(\tau) \sum_{\alpha} \|u_{\alpha}(\tau, \cdot)\|_{H^s(S_{\alpha})}^2 d\tau \right)^{\frac{1}{2}} \left(\int_0^t \chi(\tau)^{-1/2} d\tau \right)^{\frac{1}{2}} \\ &\leq \sum_a D_{\alpha 0}^{\frac{3}{2}-s} + C_2 t^{\frac{1-\gamma\theta}{2}}, \quad C_2 = (C_1 \delta \mathcal{E}(0))^{\frac{1}{2}} \left(\int_0^{\infty} \chi(\tau)^{-1/2} d\tau \right)^{\frac{1}{2}}. \end{aligned}$$

Hence, we end up with the improved dispersion bound

$$D(t) \leq C'_D (1+t)^{\gamma'}, \quad \gamma' = \frac{1-\theta\gamma}{3-2s}.$$

This argument can be repeated: since $\theta\gamma_s = \frac{\theta}{3-2s} < 1$, the mapping $\gamma \rightarrow \gamma'$ is contraction and hence the iterations converge to the unique fixed point, $\gamma_{\infty} = \frac{1}{3-2s+\theta}$. In particular, after finitely many iterations we will reach the improved rate, $\gamma_{\epsilon} = \frac{1+\epsilon}{3-2s+\theta}$, $\epsilon > 0$,

$$D(t) \leq C_{D,\epsilon} (1+t)^{\gamma_{\epsilon}}, \quad \gamma_{\epsilon} = \frac{1+\epsilon}{3-2s+\theta}.$$

Here $\epsilon > 0$ can be arbitrarily small. □

Bibliography

- [1] Yohanna Arboleda-Estudillo, Michael Krieg, Jan Stühmer, Nicholas A Licata, Daniel J Muller, and Carl-Philipp Heisenberg. Movement directionality in collective migration of germ layer progenitors. *Current Biology*, 20(2):161–169, 2010.
- [2] Ado Adamou Abba Ari, Abdelhak Gueroui, Chafiq Titouna, Ousmane Thiare, and Zibouda Aliouat. Resource allocation scheme for 5g c-ran: A swarm intelligence based approach. *Computer Networks*, 165:106957, 2019.
- [3] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.
- [4] Iztok Lebar Bajec and Frank H Heppner. Organized flight in birds. *Animal Behaviour*, 78(4):777–789, 2009.
- [5] Gianluca Baldassarre, Domenico Parisi, and Stefano Nolfi. Distributed coordination of simulated robots based on self-organization. *Artificial life*, 12(3):289–311, 2006.
- [6] Michele Ballerini, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, Irene Giardina, Vivien Lecomte, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, et al. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the national academy of sciences*, 105(4):1232–1237, 2008.
- [7] Michele Ballerini, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, Irene Giardina, Vivien Lecomte, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, et al. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the national academy of sciences*, 105(4):1232–1237, 2008.
- [8] Karla Ballman. Large deviations, techniques, and applications. *The American Mathematical Monthly*, 105(9):884, 1998.
- [9] Ch Becco, Nicolas Vandewalle, Johann Delcourt, and Pascal Poncin. Experimental evidences of a structural and dynamical transition in fish school. *Physica A: Statistical Mechanics and its Applications*, 367:487–493, 2006.
- [10] Julio M Belmonte, Gilberto L Thomas, Leonardo G Brunnet, Rita MC de Almeida, and Hugues Chaté. Self-propelled particle model for cell-sorting phenomena. *Physical Review Letters*, 100(24):248702, 2008.

- [11] Dario Benedetto, Emanuele Caglioti, and Mario Pulvirenti. A kinetic equation for granular media. *ESAIM: Mathematical Modelling and Numerical Analysis*, 31(5):615–641, 1997.
- [12] Dimitri Bertsekas. *Nonlinear Programming*, volume 4. Athena Scientific, 2016.
- [13] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.
- [14] Jerome Buhl, David JT Sumpter, Iain D Couzin, Joe J Hale, Emma Despland, Edgar R Miller, and Steve J Simpson. From disorder to order in marching locusts. *Science*, 312(5778):1402–1406, 2006.
- [15] Tariq Butt, Tabish Mufti, Ahmad Humayun, Peter B Rosenthal, Sohaib Khan, Shahid Khan, and Justin E Molloy. Myosin motors drive long range alignment of actin filaments 2. *Journal of Biological Chemistry*, 285(7):4964–4974, 2010.
- [16] José A Canizo, José A Carrillo, and Jesús Rosado. A well-posedness theory in measures for some kinetic models of collective motion. *Mathematical Models and Methods in Applied Sciences*, 21(03):515–539, 2011.
- [17] José A Carrillo, Young-Pil Choi, Piotr B Mucha, and Jan Peszek. Sharp conditions to avoid collisions in singular cucker–smale interactions. *Nonlinear Analysis: Real World Applications*, 37:317–328, 2017.
- [18] José A Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. An analytical framework for consensus-based global optimization method. *Mathematical Models and Methods in Applied Sciences*, 28(06):1037–1066, 2018.
- [19] José A Carrillo, Massimo Fornasier, Giuseppe Toscani, and Francesco Vecil. Particle, kinetic, and hydrodynamic models of swarming. In *Mathematical modeling of collective behavior in socio-economic and life sciences*, pages 297–336. Springer, 2010.
- [20] José A Carrillo, Shi Jin, Lei Li, and Yuhua Zhu. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 27:S5, 2021.
- [21] Tanmay Chakraborty and Soumya Kanti Datta. Application of swarm intelligence in internet of things. In *2017 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 67–68. IEEE, 2017.
- [22] Gui Qiang Chen, Qiang Du, and Eitan Tadmor. Spectral viscosity approximations to multidimensional scalar conservation laws. *Mathematics of Computation*, 61(204):629–643, 1993.
- [23] Shi Cheng, Lantian Guo, Tao Yang, Jiqiang Feng, Yifei Sun, Chang Shao, and Qiqi Duan. Cloud service resource allocation with particle swarm optimization algorithm. In *International Conference on Bio-Inspired Computing: Theories and Applications*, pages 523–532. Springer, 2017.

- [24] Shi Cheng, Yuhui Shi, Quande Qin, and Ruibin Bai. Swarm intelligence in big data analytics. In *International conference on intelligent data engineering and automated learning*, pages 417–426. Springer, 2013.
- [25] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.
- [26] Luis H Cisneros, Ricardo Cortez, Christopher Dombrowski, Raymond E Goldstein, and John O Kessler. Fluid dynamics of self-propelled microorganisms, from individuals to concentrated populations. In *Animal Locomotion*, pages 99–115. Springer, 2010.
- [27] Jack Coughlin and Jingwei Hu. Efficient dynamical low-rank approximation for the vlasov-ampère-fokker-planck system. *arXiv preprint arXiv:2204.10896*, 2022.
- [28] Iain D Couzin and Nigel R Franks. Self-organized lane formation and optimized traffic flow in army ants. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1511):139–146, 2003.
- [29] Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on automatic control*, 52(5):852–862, 2007.
- [30] Felipe Cucker and Steve Smale. On the mathematics of emergence. *Japanese Journal of Mathematics*, 2(1):197–227, 2007.
- [31] András Czirók, Eshel Ben-Jacob, Inon Cohen, and Tamás Vicsek. Formation of complex bacterial colonies via self-generated vortices. *Physical Review E*, 54(2):1791, 1996.
- [32] Andras Czirók, Mitsugu Matsushita, and Tamas Vicsek. Theory of periodic swarming of bacteria: Application to proteus mirabilis. *Physical Review E*, 63(3):031915, 2001.
- [33] Javier de Frutos, Volker John, and Julia Novo. Projection methods for incompressible flow problems with weno finite difference schemes. *Journal of Computational Physics*, 309:368–386, 2016.
- [34] Eleonora Di Nezza, Giampiero Palatucci, and Enrico Valdinoci. Hitchhikers guide to the fractional sobolev spaces. *Bulletin des sciences mathématiques*, 136(5):521–573, 2012.
- [35] Tam Do, Alexander Kiselev, Lenya Ryzhik, and Changhui Tan. Global regularity for the fractional euler alignment system. *Archive for Rational Mechanics and Analysis*, 228(1):1–37, 2018.
- [36] Frederick Ducatelle, Gianni A Di Caro, Alexander Förster, Michael Bonani, Marco Dorigo, Stéphane Magnenat, Francesco Mondada, Rehan O’Grady, Carlo Pinciroli, Philippe Régnier, et al. Cooperative navigation in robotic swarms. *Swarm Intelligence*, 8(1):1–33, 2014.
- [37] Bernd Einfeldt. On godunov-type methods for gas dynamics. *SIAM Journal on numerical analysis*, 25(2):294–318, 1988.

- [38] Lukas Einkemmer, Jingwei Hu, and Lexing Ying. An efficient dynamical low-rank algorithm for the boltzmann-bgk equation close to the compressible viscous flow regime. *SIAM Journal on Scientific Computing*, 43(5):B1057–B1080, 2021.
- [39] Björn Engquist and Stanley Osher. One-sided difference approximations for nonlinear conservation laws. *Mathematics of Computation*, 36(154):321–351, 1981.
- [40] Jolyon J Faria, John RG Dyer, Colin R Tosh, and Jens Krause. Leadership and social information use in human crowds. *Animal Behaviour*, 79(4):895–901, 2010.
- [41] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.
- [42] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak mathematical journal*, 25(4):619–633, 1975.
- [43] Miroslav Fiedler. Laplacian of graphs and algebraic connectivity. *Banach Center Publications*, 1(25):57–70, 1989.
- [44] Alessio Figalli and Moon-Jin Kang. A rigorous derivation from the kinetic cucker–smale model to the pressureless euler system with nonlocal alignment. *Analysis & PDE*, 12(3):843–866, 2018.
- [45] Peter Friedl and Darren Gilmour. Collective cell migration in morphogenesis, regeneration and cancer. *Nature reviews Molecular cell biology*, 10(7):445–457, 2009.
- [46] Jennifer Galanis, Daniel Harries, Dan L Sackett, Wolfgang Losert, and Ralph Nossal. Spontaneous patterning of confined granular rods. *Physical review letters*, 96(2):028002, 2006.
- [47] Simon Garnier, Christian Jost, Raphaël Jeanson, Jacques Gautrais, Masoud Asadpour, Gilles Caprari, and Guy Theraulaz. Aggregation behaviour as a source of collective decision in a group of cockroach-like-robots. In *European conference on artificial life*, pages 169–178. Springer, 2005.
- [48] Sergei Godunov and I Bohachevsky. Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics. *Matematičeskij sbornik*, 47(3):271–306, 1959.
- [49] Sigal Gottlieb, David I Ketcheson, and Chi-Wang Shu. *Strong stability preserving Runge-Kutta and multistep time discretizations*. World Scientific, 2011.
- [50] Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM review*, 43(1):89–112, 2001.
- [51] Seung-Yeal Ha and Eitan Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *arXiv preprint arXiv:0806.2182*, 2008.
- [52] Ami Harten. High resolution schemes for hyperbolic conservation laws. *Journal of computational physics*, 135(2):260–278, 1997.

- [53] Ami Harten and Stanley Osher. Uniformly high-order accurate nonoscillatory schemes. i. In *Upwind and High-Resolution Schemes*, pages 187–217. Springer, 1997.
- [54] Amiram Harten, James M Hyman, Peter D Lax, and Barbara Keyfitz. On finite-difference approximations and entropy conditions for shocks. *Communications on pure and applied mathematics*, 29(3):297–322, 1976.
- [55] Amiram Harten, Peter D Lax, and Bram van Leer. On upstream differencing and godunov-type schemes for hyperbolic conservation laws. *SIAM review*, 25(1):35–61, 1983.
- [56] Y Hayakawa. Spatiotemporal dynamics of skeins of wild geese. *EPL (Europhysics Letters)*, 89(4):48004, 2010.
- [57] Siming He and Eitan Tadmor. Global regularity of two-dimensional flocking hydrodynamics. *Comptes Rendus Mathématique*, 355(7):795–805, 2017.
- [58] Siming He and Eitan Tadmor. A game of alignment: collective behavior of multi-species. In *Annales de l’Institut Henri Poincaré C, Analyse non linéaire*, volume 38, pages 1031–1053. Elsevier, 2021.
- [59] Dirk Helbing, Illés Farkas, and Tamas Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490, 2000.
- [60] Dirk Helbing, Frank Schweitzer, Joachim Keltsch, and Peter Molnar. Active walker model for the formation of human and animal trail systems. *Physical review E*, 56(3):2527, 1997.
- [61] Charlotte K Hemelrijk and Hanspeter Kunz. Density distribution and size sorting in fish schools: an individual-based model. *Behavioral Ecology*, 16(1):178–187, 2005.
- [62] Jan Hesthaven and Sidi Mahmoud Kaber. Padé-jacobi approximants. 2016.
- [63] Jan S Hesthaven, Sigal Gottlieb, and David Gottlieb. *Spectral methods for time-dependent problems*, volume 21. Cambridge University Press, 2007.
- [64] John H Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992.
- [65] Juntao Huang, Yingda Cheng, Andrew J Christlieb, and Luke F Roberts. Machine learning moment closure models for the radiative transfer equation i: directly learning a gradient based closure. *Journal of Computational Physics*, 453:110941, 2022.
- [66] Juntao Huang, Yingda Cheng, Andrew J Christlieb, and Luke F Roberts. Machine learning moment closure models for the radiative transfer equation iii: enforcing hyperbolicity and physical characteristic speeds. *Journal of Scientific Computing*, 94(1):1–27, 2023.
- [67] Juntao Huang, Yingda Cheng, Andrew J Christlieb, Luke F Roberts, and Wen-An Yong. Machine learning moment closure models for the radiative transfer equation ii: enforcing global hyperbolicity in gradient based closures. *arXiv preprint arXiv:2105.14410*, 2021.

- [68] Michael Ibele, Thomas E Mallouk, and Ayusman Sen. Schooling behavior of light-powered autonomous micromotors in water. *Angewandte Chemie*, 121(18):3358–3362, 2009.
- [69] Zdzislaw Jackiewicz and R Park. A strategy for choosing gegenbauer reconstruction parameters for numerical stability. *Applied mathematics and computation*, 212(2):418–434, 2009.
- [70] Momin Jamil and Xin-She Yang. A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2):150–194, 2013.
- [71] Guang-Shan Jiang and Eitan Tadmor. Nonoscillatory central schemes for multidimensional hyperbolic conservation laws. *SIAM Journal on Scientific Computing*, 19(6):1892–1917, 1998.
- [72] Moon-Jin Kang and Alexis F Vasseur. Asymptotic analysis of vlasov-type equations under strong local alignment regime. *Mathematical Models and Methods in Applied Sciences*, 25(11):2153–2173, 2015.
- [73] Dervis Karaboga et al. An idea based on honey bee swarm for numerical optimization. Technical report, Technical report-tr06, Erciyes university, engineering faculty, computer . . . , 2005.
- [74] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.
- [75] Trygve K Karper, Antoine Mellet, and Konstantina Trivisa. Hydrodynamic limit of the kinetic cucker–smale flocking model. *Mathematical Models and Methods in Applied Sciences*, 25(01):131–163, 2015.
- [76] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN’95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE, 1995.
- [77] David A Kessler and Herbert Levine. Pattern formation in dictyostelium via the dynamics of cooperative biological entities. *Physical Review E*, 48(6):4801, 1993.
- [78] Taher Khadhraoui, Sami Ktata, Faouzi Benzarti, and Hamid Amiri. Features selection based on modified pso algorithm for 2d face recognition. In *2016 13th international conference on computer graphics, imaging and visualization (CGiV)*, pages 99–104. IEEE, 2016.
- [79] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [80] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.
- [81] Othmar Koch and Christian Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):434–454, 2007.
- [82] Kaipa N Krishnanand and Debasish Ghose. Detection of multiple source locations using a glowworm metaphor with applications to collective robotics. In *Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005.*, pages 84–91. IEEE, 2005.
- [83] Arshad Kudrolli. Concentration dependent diffusion of self-propelled rods. *Physical Review Letters*, 104(8):088001, 2010.
- [84] Arshad Kudrolli, Geoffroy Lumay, Dmitri Volfson, and Lev S Tsimring. Swarming and swirling in self-propelled polar granular rods. *Physical review letters*, 100(5):058001, 2008.
- [85] Alexander Kurganov and Doron Levy. A third-order semidiscrete central scheme for conservation laws and convection-diffusion equations. *SIAM Journal on Scientific Computing*, 22(4):1461–1488, 2000.
- [86] Alexander Kurganov, Sebastian Noelle, and Guergana Petrova. Semidiscrete central-upwind schemes for hyperbolic conservation laws and hamilton–jacobi equations. *SIAM Journal on Scientific Computing*, 23(3):707–740, 2001.
- [87] Alexander Kurganov and Eitan Tadmor. New high-resolution central schemes for nonlinear conservation laws and convection–diffusion equations. *Journal of Computational Physics*, 160(1):241–282, 2000.
- [88] Stanis law Łojasiewicz. Ensembles semi-analytiques. *IHES notes*, 1965.
- [89] Peter D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Communications on Pure and Applied Mathematics*, 7:159–193, 1954.
- [90] Doron Levy, Gabriella Puppo, and Giovanni Russo. Central weno schemes for hyperbolic systems of conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(3):547–571, 1999.
- [91] Doron Levy, Gabriella Puppo, and Giovanni Russo. A third order central weno scheme for 2d conservation laws. *Applied Numerical Mathematics*, 33(1-4):415–421, 2000.
- [92] Doron Levy, Gabriella Puppo, and Giovanni Russo. A fourth-order central weno scheme for multidimensional hyperbolic systems of conservation laws. *SIAM Journal on scientific computing*, 24(2):480–506, 2002.
- [93] Xu-Dong Liu and Eitan Tadmor. Third order nonoscillatory central scheme for hyperbolic conservation laws. *Numerische mathematik*, 79(3):397–425, 1998.
- [94] Stanislas Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l’institut Fourier*, volume 43, pages 1575–1595, 1993.

- [95] Jingcheng Lu and Eitan Tadmor. Hydrodynamic alignment with pressure ii. multi-species. *Quarterly of Applied Mathematics*, 2022.
- [96] Jingcheng Lu, Eitan Tadmor, and Anil Zenginoglu. Swarm-based gradient descent method for non-convex optimization. *arXiv preprint arXiv:2211.17157*, 2022.
- [97] Yonghe Lu, Minghui Liang, Zeyuan Ye, and Lichao Cao. Improved particle swarm optimization algorithm and its application in text feature selection. *Applied Soft Computing*, 35:629–636, 2015.
- [98] M Min, SM Kaber, and W Don. Fourier–padé approximations and filtering for spectral simulations of an incompressible boussinesq convection problem. *Mathematics of computation*, 76(259):1275–1290, 2007.
- [99] Piotr Minakowski, Piotr B Mucha, Jan Peszek, and Ewelina Zatorska. Singular cuckoo–smale dynamics. In *Active Particles, Volume 2*, pages 201–243. Springer, 2019.
- [100] Sebastien Motsch and Eitan Tadmor. A new model for self-organized dynamics and its flocking behavior. *Journal of Statistical Physics*, 144(5):923–947, 2011.
- [101] Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM review*, 56(4):577–621, 2014.
- [102] Wim A Mulder and Bram Van Leer. Experiments with implicit upwind methods for the euler equations. *Journal of Computational Physics*, 59(2):232–246, 1985.
- [103] Seido Nagano. Diffusion-assisted aggregation and synchronization in dictyostelium discoideum. *Physical review letters*, 80(21):4826, 1998.
- [104] Vijay Narayan, Narayanan Menon, and Sriram Ramaswamy. Nonequilibrium steady states in a vibrated-rod monolayer: tetratic, nematic, and smectic correlations. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(01):P01005, 2006.
- [105] Vijay Narayan, Sriram Ramaswamy, and Narayanan Menon. Long-lived giant number fluctuations in a swarming granular nematic. *Science*, 317(5834):105–108, 2007.
- [106] Roberto Natalini and Thierry Paul. On the mean field limit for cuckoo-smale models. *arXiv preprint arXiv:2011.12584*, 2020.
- [107] Haim Nessyahu and Eitan Tadmor. Non-oscillatory central differencing for hyperbolic conservation laws. *Journal of computational physics*, 87(2):408–463, 1990.
- [108] Shervin Nouyan, Roderich Gross, Michael Bonani, Francesco Mondada, and Marco Dorigo. Teamwork in self-organized robot colonies. *IEEE Transactions on Evolutionary Computation*, 13(4):695–711, 2009.
- [109] Hailiang Liu null and Xuping Tian. An adaptive gradient method with energy and momentum. *Annals of Applied Mathematics*, 38(2):183–222, jun 2022.

- [110] Rehan O’Grady, Roderich Groß, Anders Lyhne Christensen, and Marco Dorigo. Self-assembly strategies in a group of autonomous mobile robots. *Autonomous Robots*, 28(4) : 439 – 455, 2010.
- [111] Stanley Osher. Riemann solvers, the entropy condition, and difference. *SIAM Journal on Numerical Analysis*, 21(2):217–235, 1984.
- [112] Stanley Osher. Convergence of generalized muscl schemes. *SIAM Journal on Numerical Analysis*, 22(5):947–961, 1985.
- [113] Julia K Parrish, Steven V Viscido, and Daniel Grunbaum. Self-organized fish schools: an examination of emergent properties. *The biological bulletin*, 202(3):296–305, 2002.
- [114] Jan Peszek. Discrete cucker–smale flocking model with a weakly singular weight. *SIAM Journal on Mathematical Analysis*, 47(5):3671–3686, 2015.
- [115] Giovanni Pini, Arne Brutschy, Marco Frison, Andrea Roli, Marco Dorigo, and Mauro Birattari. Task partitioning in swarms of robots: An adaptive method for strategy selection. *Swarm Intelligence*, 5(3):283–304, 2011.
- [116] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 27(01):183–204, 2017.
- [117] Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17–32, 1964.
- [118] David Poyato and Juan Soler. Euler-type equations and commutators in singular and hyperbolic limits of kinetic cucker–smale models. *Mathematical Models and Methods in Applied Sciences*, 27(06):1089–1152, 2017.
- [119] Wouter-Jan Rappel, Alastair Nicol, Armand Sarkissian, Herbert Levine, and William F Loomis. Self-organized vortex state in two-dimensional dictyostelium dynamics. *Physical review letters*, 83(6):1247, 1999.
- [120] Philip L Roe. Approximate riemann solvers, parameter vectors, and difference schemes. *Journal of computational physics*, 43(2):357–372, 1981.
- [121] Richard Saurel and Rémi Abgrall. A simple method for compressible multifluid flows. *SIAM Journal on Scientific Computing*, 21(3):1115–1145, 1999.
- [122] Volker Schaller, Christoph Weber, Christine Semmrich, Erwin Frey, and Andreas R Bausch. Polar patterns of driven filaments. *Nature*, 467(7311):73–77, 2010.
- [123] Chi-Wang Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In *Advanced numerical approximation of non-linear hyperbolic equations*, pages 325–432. Springer, 1998.

- [124] Michael L Shur, Philippe R Spalart, and Michael Kh Strelets. Noise prediction for increasingly complex jets. part i: Methods and tests. *International journal of aeroacoustics*, 4(3):213–245, 2005.
- [125] Roman Shvydkoy and Eitan Tadmor. Eulerian dynamics with a commutator forcing. *Transactions of Mathematics and its Applications*, 1(1), 2017.
- [126] Roman Shvydkoy and Eitan Tadmor. Eulerian dynamics with a commutator forcing ii: Flocking. *arXiv preprint arXiv:1701.07710*, 2017.
- [127] Roman Shvydkoy and Eitan Tadmor. Eulerian dynamics with a commutator forcing iii. fractional diffusion of order $0 < \alpha < 1$. *Physica D: Nonlinear Phenomena*, 376:131–137, 2018.
- [128] Roman Shvydkoy and Eitan Tadmor. Topologically based fractional diffusion and emergent dynamics with short-range interactions. *SIAM Journal on Mathematical Analysis*, 52(6):5792–5839, 2020.
- [129] Andrey Sokolov, Igor S Aranson, John O Kessler, and Raymond E Goldstein. Concentration dependence of the collective dynamics of swimming bacteria. *Physical review letters*, 98(15):158102, 2007.
- [130] Onur Soysal and Erol Sahin. Probabilistic aggregation strategies in swarm robotic systems. In *Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005.*, pages 325–332. IEEE, 2005.
- [131] William M Spears, Diana F Spears, Jerry C Hamann, and Rodney Heil. Distributed, physics-based control of swarms of vehicles. *Autonomous robots*, 17(2):137–162, 2004.
- [132] Raymond J Spiteri and Steven J Ruuth. A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM Journal on Numerical Analysis*, 40(2):469–491, 2002.
- [133] Rainer Storn and Kenneth Price. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.
- [134] Michael Struwe and M Struwe. *Variational methods*, volume 991. Springer, 2000.
- [135] Nobuhiko J Suematsu, Satoshi Nakata, Akinori Awazu, and Hiraku Nishimori. Collective behavior of inanimate boats. *Physical Review E*, 81(5):056210, 2010.
- [136] A Suresh and HT Huynh. Accurate monotonicity-preserving schemes with runge–kutta time stepping. *Journal of Computational Physics*, 136(1):83–99, 1997.
- [137] Peter K Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM journal on numerical analysis*, 21(5):995–1011, 1984.

- [138] Balint Szabo, GJ Szöllösi, B Gönci, Zs Jurányi, David Selmeczi, and Tamás Vicsek. Phase transition in the collective migration of tissue cells: experiment and model. *Physical Review E*, 74(6):061908, 2006.
- [139] Eitan Tadmor. The large-time behavior of the scalar, genuinely nonlinear lax-friedrichs scheme. *Mathematics of computation*, 43(168):353–368, 1984.
- [140] Eitan Tadmor. Convenient total variation diminishing conditions for nonlinear difference schemes. *SIAM journal on numerical analysis*, 25(5):1002–1014, 1988.
- [141] Eitan Tadmor. On the mathematics of swarming: emergent behavior in alignment dynamics. *arXiv preprint arXiv:2102.09134*, 2021.
- [142] Eitan Tadmor. Swarming: hydrodynamic alignment with pressure. *Bull. of AMS, accepted, arXiv preprint arXiv:2208.11786*, 2022.
- [143] Eitan Tadmor and Changhui Tan. Critical thresholds in flocking hydrodynamics with non-local alignment. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2028):20130401, 2014.
- [144] John Toner and Yuhai Tu. Flocks, herds, and schools: A quantitative theory of flocking. *Physical review E*, 58(4):4828, 1998.
- [145] Eleuterio F Toro, Michael Spruce, and William Speares. Restoration of the contact surface in the hll-riemann solver. *Shock waves*, 4(1):25–34, 1994.
- [146] Ali E Turgut, Hande Çelikkanat, Fatih Gökçe, and Erol Şahin. Self-organized flocking in mobile robot swarms. *Swarm Intelligence*, 2(2):97–120, 2008.
- [147] Gale Dick Van Albada, B van Leer, and WWjun Roberts. A comparative study of computational methods in cosmic gas dynamics. In *Upwind and high-resolution schemes*, pages 95–103. Springer, 1997.
- [148] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.
- [149] Bram Van Leer. On the relation between the upwind-differencing schemes of godunov, engquist–osher and roe. *SIAM Journal on Scientific and statistical Computing*, 5(1):1–20, 1984.
- [150] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.
- [151] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.

- [152] Ashley JW Ward, David JT Sumpter, Iain D Couzin, Paul JB Hart, and Jens Krause. Quorum decision-making facilitates information transfer in fish shoals. *Proceedings of the National Academy of Sciences*, 105(19):6948–6953, 2008.
- [153] Justin K Werfel, Kirsten Petersen, and Radhika Nagpal. Distributed multi-robot algorithms for the termite 3d collective construction system. In *Proceedings of Robotics: Science and Systems*. Institute of Electrical and Electronics Engineers, 2011.
- [154] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in neural information processing systems*, 30, 2017.
- [155] Philip Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.
- [156] Paul Woodward and Phillip Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. *Journal of computational physics*, 54(1):115–173, 1984.
- [157] Xin-She Yang. *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.
- [158] Ju Zhang and Thomas L Jackson. A high-order incompressible flow solver with weno. *Journal of Computational Physics*, 228(7):2426–2442, 2009.