Models for Pattern Formation in Biological Systems

by

Nastassia Pouradier Duteil

A dissertation submitted to the

Graduate School - Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

Written under the direction of

Benedetto Piccoli

And approved by

Benedetto Piccoli

Nir Yakoby

Emmanuel Trélat

Siqi Fu

Camden, New Jersey

May 2017

THESIS ABSTRACT

Models for Pattern Formation in Biological Systems

by NASTASSIA POURADIER DUTEIL

Thesis Director:
Benedetto Piccoli

This thesis presents models for two different kinds of pattern formation in biological systems. Developmental patterning refers to pattern formation by cell differentiation during an organism's development. Cell differentiation is controlled by morphogens, signaling molecules that diffuse in a growing organism. We focus on the specific case of the activation of the epidermal growth factor receptor pathway, a highly-conserved signaling pathway across animals, that controls both the posterior-anterior and the dorso-ventral axes during *Drosophila* oogenesis. Not only can the diffusion of morphogens control the growth of an organism, but the diffusion itself is influenced by the changing geometry of the domain. We develop a mathematical framework enabling this double coupling of the diffusion of a signal on a time-evolving Riemannian manifold and the evolution of the manifold via a vector field depending on the diffusing signal: "Developmental Partial Differential Equations".

The second kind of pattern formation, behavioral patterning, arises from local interactions between individuals that lead to a global group behavior. We focus on several of these models, also referred to as Social Dynamics systems. We examine how to control the dynamics to guide the system to a target configuration. In particular, we study the optimal control of a collective migration model to guide the system to consensus at a target velocity, as well as the controllability away from consensus of an opinion dynamics system. We analyze the influence of the state space on the dynamics by designing a general opinion dynamics model on Riemannian manifolds. Lastly, we investigate the role of the interaction network on the periodicity of the dynamics, specifically in creating a "social choreography".

A Claire, Xavier, Dimitri, Vladimir et Basile.

Acknowledgements

This work would not have been possible without the guidance of my advisor Dr. Benedetto Piccoli. Thank you for your time, your valuable advice, and your dedication, as well as for your warm reception when I first arrived, your impromptu and captivating lessons to the lab on Fridays, and your patience with my never-ending questions. Thank you also to Dr. Nir Yakoby, for hosting me at the Waterfront Technology Center, and for your always enthusiastic explanations of biological mechanisms, not scientifically valid unless drawn on lab napkins or on the windows. I never thought that I would ever meet somebody so passionate about flies! Thank you to Dr. Emmanuel Trélat for taking the time to work with me during my visits to Paris. Our interactions have always been pleasant and productive. I value the scientific and academic guidance you have given me, and I look forward to continuing working with you in the future. Thank you also to Dr. Siqi Fu for accepting to be on my PhD committee. I greatly admire the excellence of your work and your commitment to Rutgers-Camden.

Among all my collaborators, I wish to thank particularly warmly Dr. Francesco Rossi. Thank you for our many fruitful interactions, your incredible hospitality in Marseille, our many Skype meetings, and your passionate approach to research which is a great source of inspiration for me. Thank you of course to Nicole Revaitis for the quality of your work and your contribution in making this bio-mathematical collaboration harmonious. Thank you to Matthew Niepielko, to Benjamin Scharf, to Michael Herty, to Jingmei Qiu.

Thank you to my former and current fellow lab members for their support. Thanks to all my Rutgers friends for supporting me and helping me grow as a person and as a scientist: Dan, Steve, Maria Laura, Aylin, Sung, Gina, Stefán, Ruchi, Sruthi, Lekha, Min, Ammar, Slim, Kuhn, Jen... Nandri da Harish. I also wish to thank all the friends who checked on me and supported me from afar or by visiting, whether from France, Russia, Japan, Australia or Canada.

And lastly, and most importantly, my loving family. Merci à mes grands-parents qui avec une patience infinie me demandent seulement une fois par mois quand je rentre en France. Merci à tous mes oncles, tantes et cousins qui trouvent que le vol migratoire des oiseaux est bien plus poétique que le développement des drosophiles. Merci à mes frères d'avoir vérifié ligne par ligne la rigueur de ce travail, et pour leur intérêt particulier pour l'esthétisme... Et merci à mes parents d'être là envers et contre tout, malgré la distance.

Contents

Introduction

There is a long history of collaboration between biologists and mathematicians, but recently the interdisciplinary field of Systems Biology has seen an even greater burst of interest. This is due, among other things, to advances in technology revealing the complexity of biological systems, to an increase in computing power allowing to simulate with more and more accuracy very complex systems, and to the new availability of data-rich information sets, difficult to interpret without analytic tools.

Biological systems exhibit complex behaviors, and can naturally lead to fascinating patterns. For example, in developmental biology, pattern formation refers to the generation of complex organization of cell fates in space and time. Phenotypic structures such as stripes or spots can be explained by the interaction and diffusion of morphogens in the developing organism, a mechanism known as Turing patterning [125].

On a larger scale, patterns are found in animal behavior. Groups of autonomous agents like animals exhibit strong coordination in their movements, which also leads to the creation of patterns (some examples include lines of ants, murmuration in flocks of starlings, collective evasion in schools of fish, etc.) [8, 24, 88, 89, 103, 124, 128]. In this case, the system's global behavior emerges from local interactions between individuals, a phenomenon referred to as self-organization.

The work presented in this thesis provides mathematical frameworks to investigate pattern formation in two classes of systems, respectively describing **developmental patterning** and **behavioral patterning**. These two classes of systems differ in their biological applications and in their mathematical formulations. Therefore, this thesis is divided into two parts.

One of the first successful attempts at explaining patterning in developing organisms was presented in a seminal paper by Alan Turing in 1952 [125]. Turing showed that spatial heterogeneities can arise from the reaction and diffusion of several competing chemical substances (known as morphogens). Since then, his reaction-diffusion model was developed and exploited to justify periodic patterning in various organisms, such as in the marine angelfish [64], or in digit formation during limb development [78]. In early applications of Turing's reaction-diffusion model, the mechanisms of reaction and diffusion were considered to happen on a much shorter time scale than that of the organism's growth, and the domain was thus modeled with a constant size. However, it has recently been suggested that domain growth has a non-negligible effect on the diffusion of morphogens - hence on the patterning of the organism (see among others the works of Crampin et al. [27], Kondo et al.

[64], Baker, Maini, Miura, Plaza [78, 98]). More specifically, growth of the domain has been shown to increase the robustness of pattern formation [27], a characteristic of Turing's model that had long been lacking due to its extreme sensitivity to parameter choice and initial conditions [27, 76, 98]. Plaza et al. [98] analyzed numerically the effects of linear, exponential and saturated growth on domain patterning in one dimension, as well as on a two-dimensional conic surface embedded in $\mathbb{R}^3$. In addition to growth, curvature has also been shown to impact the reaction-diffusion dynamics. In particular, Turing patterning has been studied on a sphere by Varea et al. [126, 127] and on a cone by Plaza et al. [98]. It is thus clear that growth has a non-negligible effect on the patterns formed by morphogens, as it expands the domain and changes its curvature. Conversely, several authors studied the interplay between growth and morphogen concentration by considering growth to be driven by the local morphogens concentration. Lefèvre and Mangin were able to build a model reproducing the progressive folding of the cortical surface during brain development by locally deforming the surface based on the distributions of two morphogens (an activator and an inhibitor) [71]. Harrison et al. showed how simultaneous growth, reaction and diffusion in an originally hemispherical domain are able to trigger sequential dichotomous branchings, mimicking the growth of plants [52].

Inspired by the evidence that growth and diffusion are intrinsically linked during development, we have sought to design a complete mathematical framework incorporating growth of the organism, diffusion and reaction of signals on its surface, and relative cell movements. This was done in two parallel lines of research.

In a first project, we constructed a detailed model specifically for the dynamics of the epidermal growth factor receptor (EGFR) activation in *Drosophila* oogenesis. This work emerged from a collaboration between the Piccoli laboratory of Applied Mathematics and the Yakoby laboratory of Developmental Biology of the Center for Computational and Integrative Biology at Rutgers University. The Yakoby laboratory studies the mechanisms underlying cell fate determination by cell signaling in *Drosophila* ovaries. In particular, it focuses on the EGFR signaling pathway. The EGFR is a transmembrane protein that is activated by the binding of its specific ligands. The EGFR signaling pathway is a highly-conserved pathway, active in humans as well as in *Drosophila*. It controls cell processes such as cell migration and apoptosis [20, 105]. It has been shown that mutations leading to overexpression of EGFR are associated to the development of a numerous cancers [77]. *Drosophila* oogenesis (i.e. egg formation) provides a good model system to study the functioning of the EGFR signaling pathway. During the early stages of oogenesis, EGFR signaling has been shown to be regulated by the TGFα-like ligand Gurken [86]. Mathematically, the diffusion of the Gurken protein and its interaction with other components of the pathway was modeled by the classic

coupling of a PDE (for diffusion) with ODEs (for reactions). The curvature of the egg chamber was taken into account by using the Laplace-Beltrami operator for the diffusion. The novelty of our approach is our focus on the growth of the egg chamber and on the shift of the overlying cells. We have developed a complete numerical tool to include all components of the model: growth of the manifold, cell movement, diffusion and reactions. Comparing the simulations with experimental data yielded convincing results, confirming the hypothesized crucial role of growth and cell movement in shaping the Gurken signal. This constitutes Chapter 1 of this thesis.

A parallel line of work has led us to develop a theoretical framework for diffusion equations on evolving manifolds, that we named *Developmental PDEs*. Although this has not been shown for Gurken, there exist morphogens that influence the growth of organisms [52, 71]. For this reason, we considered the growth vector-field to depend on the signal diffusing on the manifold. Here, we present this new framework and provide results of existence and uniqueness for the solution to this equation. We extended the definition of the Lie bracket to apply it to operators of different nature: a usual first-order vector field (for growth) and the second-order diffusion operator. We proved approximate controllability of the system, identified controls for particular final shapes, and performed numerical analyses in order to show how the control acts on the manifold. This work was published in [100, 104] and constitutes Chapter 2.

While the first part of this thesis deals with developmental patterning at the molecular level, our second axis of research focuses on behavioral patterning by social interaction. The emergence of a group's global behavior from local interaction rules is referred to as self-organization. We use the term Social Dynamics to indicate the study of such systems, with an emphasis on understanding the mechanisms leading from local rules to global phenomena, as well as identifying the resulting global pattern formation. A review of this field was published in [3].

Social dynamics models can be classified as first-order models and second-order models. In first-order models, we refer to the variables of interest as *opinions*, even though such models can describe a wide range of attributes such as positions, market shares or wealth. The opinion of each agent is affected by neighboring agents' opinions in the state space. On the other hand, in second-order models, the variables of interest are the *velocities*, obtained as the time derivatives of the positions. Each agent's velocity is affected by the velocities of agents whose positions are close in the state-space. First-order models (or opinion dynamics) can give rise to patterns such as *consensus* (i.e. agreement of all states), *polarization* (i.e. disagreement between two opposite parties) or *clustering* (i.e. break-down of the opinions into several subsets). A first formulation of opinion dynamics can be traced back to French's research on social influence [40], followed by works by Harary [51],

De Groot [33] and Lehrer [72], all focusing on linear models. More recently, nonlinear models were introduced and analyzed by Krause [66, 67], Dittmer [35], Hegselmann and Flache [54]. Second-order models are commonly applied to animal groups to study coordinated collective behavior (as done by Couzin et al. [26], Cristiani, Frasca and Piccoli [28], Giardina [43], Krause and Ruxton [65], Leonard [73] and Sumpter [122]) for example in fish (Huth and Vissel [59], Parrish, Viscido and Grunbaum [89]) or birds (Ballerini et al. and Cucker and Smale [6, 31]). Some models have been designed to include simple interaction rules like attraction, short-distance repulsion and mimetic orientation or alignment. Agreement of all agents in the velocity variable is referred to as *alignment* or *flocking*.

A large number of applications of Social Dynamics models involve the control of robotic networks or autonomous vehicles, as done by Bullo, Cortés, and Martínez [15]. Control is used to impose consensus or alignment when it is not reached naturally (see Caponigro, Fornasier, Piccoli, and Trélat [17, 18]), or to guide the agents in a specific direction, as done by Leonard for the migration of animal groups [73]. Ways of controlling the system include spreading leaders among the group or acting on the network. The work presented in Chapters 3 and 4 focuses on how to influence pattern formation by controlling the system in order to guide it to a target behavior.

In Chapter 3, we introduce a migration model inspired by the behavior of groups of migrating animals all moving towards a common destination. We designed optimal control strategies to choose leaders among the group in order to drive the group to consensus at the preassigned migration velocity. This work was published in [94].

In Chapter 4 we studied avoidance of consensus, that is how to act on a system to keep the agents as far from each other as possible. This approach aims to avoid "black swan" phenomena, which characterize rare events with a large impact and a possible retrospective justification [7, 123]. Applications involve preventing various kinds of dangerous clustering, for instance of opinions (to avoid single-party systems), of financial shares (to avoid collapse of the market), or of individuals (to avoid high crowd densities that can lead to stampedes).

The last chapter of this thesis focuses on understanding the influence of the state-space on the dynamics. Most studies have considered dynamics in Euclidean spaces (most often 1-dimensional for opinion models and 2 or 3-dimensional for animal groups). One can also study opinion dynamics on general Riemannian manifolds. For instance, Caponigro, Lai and Piccoli studied a nonlinear model of opinion formation on the sphere [19], with a rich structure leading to unusual equilibria. Consensus dynamics on special orthogonal groups were investigated, for example by Sarlette and Sepulchre [106, 107, 108], motivated by applications to satellites or ground vehicles. Chapter 5 presents a general model for opinion dynamics on Riemannian manifolds.

Part I

Reaction-diffusion equations on time-evolving manifolds

Chapter 1

Modeling the evolution of a ligand in *Drosophila*

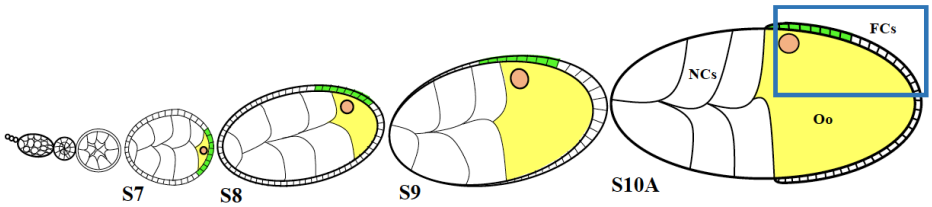
Introduction

Developmental biology is the study of how organisms grow and develop, and, in particular, of the genetic control of cell growth, differentiation and morphogenesis. Tissue patterning and cell fates are determined by cell signaling pathways, which are triggered by the binding of signaling molecules to cell receptors. These molecules are referred to as "morphogens", a term coined by Alan Turing in his groundbreaking article "The Chemical basis of morphogenesis" [125]. The complexity of signaling pathways has encouraged collaborative work between developmental biologists and applied mathematicians, with the construction of mathematical models aiming to reproduce the interactions between components of the pathways and justify the signaling patterns observed experimentally [22, 45, 101, 102].

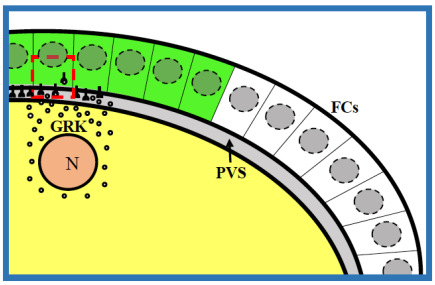
In this line of thought, a collaboration was created between the Piccoli laboratory of Applied Mathematics and the Yakoby laboratory of Developmental Biology of the Center for Computational and Integrative Biology at Rutgers University, with the aim of modeling the dynamics of the EGFR signaling pathway in the fruit fly ovaries. The Yakoby laboratory studies the mechanisms underlying cell fate determination by cell signaling in *Drosophila* species. *Drosophila* is a commonly used model organism in genetics and developmental biology. Some of the practical reasons for its wide use in research include its short life cycle, its cheap cost and its ability to survive and reproduce in large numbers in a lab environment [118]. Moreover, the genomes of several species such as *D. melanogaster, D. willistoni* and *D. virilis* have been fully sequenced [41]. Within *Drosophila* development, oogenesis is the focus of this chapter. A female *Drosophila* has 2 ovaries made of about 16 ovarioles [118]. Each ovariole contains egg chambers, the precursors of the mature eggs, at all stages of egg development, as in an assembly line (Fig. 1.1a). For the purpose of this work,

we focus on Stage 7 to Stage 10A of oogenesis. Egg chambers at these developmental stages consist of 15 nurse cells, one oocyte, and a layer of about 1000 follicle cells surrounding the developing oocyte [118]. The follicle cells are separated from the germ cells (nurse cells and oocyte) by a thin region called perivitelline space. During these four stages, the egg chamber undergoes mechanical transformations: its dimensions increase by a factor of four; the oocyte grows inside the egg chamber; the oocyte nucleus transitions from being anchored at the posterior end to the dorsal anterior position; the follicle cells shift from anterior to posterior (Fig. 1.1a).

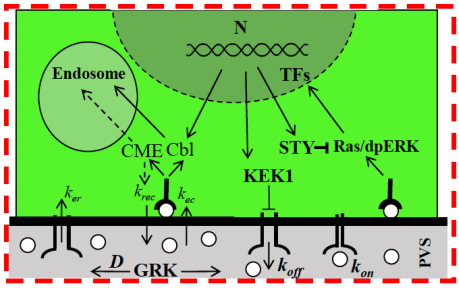
In the early stages of *Drosophila* oogenesis, EGFR signaling is activated by the binding of the TGFα-like ligand Gurken (GRK) to the EGF receptor [84, 85, 109]. This consequently sets the dorso-ventral and anterior-posterior axes of the egg chamber by altering the fates of certain cells [84, 85, 109]. Gurken is secreted from near the oocyte nucleus, diffuses in the thin perivitalline space, is internalized by EGFR and triggers a signaling cascade resulting in the double phosphorylation of ERK (extracellular signal-regulated kinases), dpERK, in the overlaying follicle cells (Fig. 1.1b and 1.1c). These interactions are combined with a physical transformation of the egg chamber: during this process, the egg chamber grows significantly and the follicle cells gradually shift from the anterior to the posterior of the egg chamber [118].



(a) Growth of the egg chamber during early stages of oogenesis and its mechanical transformations, with relative growth of the oocyte, transition of the nucleus' position from posterior to dorsal anterior, and shift of the follicle cells.



(b) Secretion of GRK from near the oocyte nucleus, diffusion of GRK in the perivitelline space and binding of GRK to the EGFR at the apical surface of the overlying follicle cells.



(c) Internalization of GRK setting off the RAS/RAF/MEK signaling cascade and the production of inhibitors that act as negative feedback.

Figure 1.1: Schematic of the mechanisms responsible for the spatio-temporal GRK evolution.

The community of biomathematicians has demonstrated the role of growth in shaping the distribution of signals in developing organisms. The effects of growth rate, curvature and cell movements on stability, geometry and growth of signaling patterns have been investigated in recent studies [2, 5, 27, 52, 64, 78, 98, 110, 126, 127]. Although the importance of domain growth has been established, there currently does not exist a complete model combining the physical transformation of the organism with its effect on the various components of a signaling pathway. More specifically, while the activation gradient of EGFR signaling was previously modeled, and it neglected the relative movement of cells and the growth of the egg chamber [45, 114, 133]. Here we present a fully integrated model of the EGFR signaling pathway during oogenesis, by taking into account not only the interaction between components of the pathway, but also the growth of the egg chamber, the relative movement of the oocyte nucleus and the shift of the overlaying follicle cells.

This chapter introduces the system studied, presents a mathematical model integrating all its components, describes the numerical challenges linked to numerical simulations, and compares numerical results with experimental data, in the overall aim of explaining how growth and cell movement contribute in shaping the distribution of the signals.

1.1 Modeling Gurken concentration in *D. melanogaster*

It has been shown that EGFR activation gives rise to cell differentiation and in particular to the formation of various structures on the *Drosophila* eggshell, among which are the dorsal appendages and the dorsal ridge [85, 86, 92]. The dorsal ridge is a lumen-like structure present along the dorsal side of eggshells in certain *Drosophila* species. This structure extends from the anterior to the posterior end of the eggshell and varies in length and width across species [86]. Niepielko and Yakoby [86] have shown that the distribution of GRK is closely linked to the size and shape of the dorsal ridge. Indeed, it was shown that the distribution of the GRK protein is consistent with the activation of EGFR, while the EGFR activation pattern is itself consistent with the length and shape of the dorsal ridge. More specifically, *D. melanogaster* does not have a dorsal ridge on its eggshell, which is consistent with the restricted activation patterns of EGFR [86]. In species with a dorsal ridge such as *D. willistoni* and *D. cardini*, the activation of EGFR localization extends further towards the posterior end.

These observations motivate the need to understand the spatio-temporal evolution of GRK distribution throughout oogenesis.

1.1.1 Reaction-diffusion model

We first present a complete reaction-diffusion model describing the spatio-temporal evolution of Gurken, EGFR and dpERK, neglecting the growth of the egg chamber and the shift of the follicle cells. These two additions to the model will be presented in Sections 1.1.2 and 1.1.3.

The EGF ligand Gurken is secreted to the perivitelline space surrounding the oocyte from near the oocyte nucleus, which is dynamically localized during oogenesis. As shown in Fig. 1.1a, the oocyte nucleus is at the posterior end at Stage 8, and later becomes anchored at the dorsal anterior of the oocyte where it remains as the oocyte grows [85, 109, 118]. The perivitelline space is an open region where the ligand can diffuse and bind to the EGF receptor (EGFR) in the overlaying follicle cells. This sets off the signaling cascade RAS/RAF/MEK. This phosphorylation cascade targets ERK which activates the molecule dpERK, which can be detected by an antibody (Figure 1.2). Then dpERK activates or deactivates transcriptional regulators, while simultaneously being inactivated via dephosphorylation or degradation by a protease (see Figure 1.1c).

We present a reaction-diffusion model that takes into account:

- The diffusion of Gurken into the perivitelline space by the moving morphogen source and its internalization;
- The time evolution of the surface receptors, receptor-ligand complexes and internalized receptors;
- The activation of dpERK through the EGFR signaling pathway;
- The action of inhibitors in a negative feedback loop.

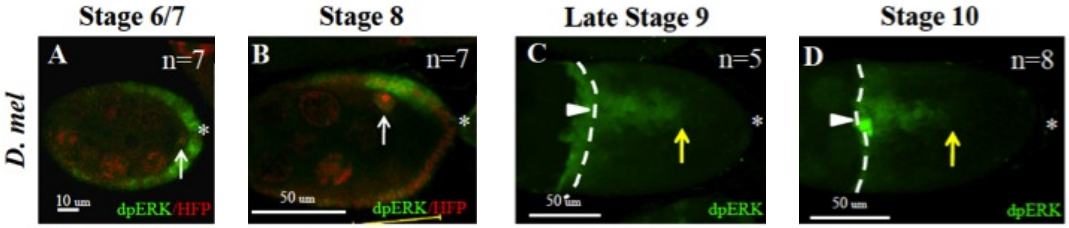


Figure 1.2: dpERK staining in stages 7 through 10A of *D. melanogaster*. Images A and B offer sagittal views and show the dynamic localization of the nucleus (indicated with an arrow) with respect to the overlaying follicle cells. Images C and D offer dorsal views of dpERK staining. The distribution of dpERK is short, in line with the absence of dorsal ridge observed in *D. melanogaster* [86].

Reactions

Let L denote the concentration of ligand. The ligand binds to the receptors that are on the apical surface of the follicle cells (FC) facing the perivitelline space (PVS). R denotes the concentration of free surface receptors, C denotes the concentration of ligand-receptor complexes and $\bar{L}$ denotes the concentration of the ligand on the surface PVS/FC.

Following [22], we also introduce the internalized receptor-ligand complexes C_{i} and the signal (dpERK) S. The surface ligand-receptor complexes C are formed when the ligand binds to receptors, with rate k_{on} . The complexes can then be internalized with rate k_{ec} , or dissociated with rate k_{off}. In turn, a fraction α_{rec} of the internalized complexes C_{i} is recycled back to the surface of the FCs with rate k_{rec} and a fraction α_{deg} is degraded with rate k_{deg} [115]. The free receptors R are produced with rate Q_r. Their concentration increases when the surface complexes C dissociate, and decreases when they bind to the ligand (with rate k_{on}) or are internalized (with rate k_{er}). From these equations we can further compute the concentration of dpERK, which is produced by the phosphorylation cascade triggered by the internalization of Gurken. The evolution of the signal S (dpERK) is characterized by the phosphorylation of the complexes via the Ras/Raf/Mek cascade (with rate k_{s} and by its degradation with rate k_{d} (encompassing the phenomena of dephosphorylation and digestion by protease). This gives the following system of ODEs, for the complexes C, internalized complexes C_{i}, receptors R and signal S:

$$\begin{cases} \dfrac{\partial C}{\partial t} = k_{\mathrm{on}} R\bar{L} - (k_{\mathrm{off}} + k_{\mathrm{ec}})C + \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} \\ \dfrac{\partial C_{\mathrm{i}}}{\partial t} = k_{\mathrm{ec}} C - \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} - \alpha_{\mathrm{deg}} k_{\mathrm{deg}} C_{\mathrm{i}} \\ \dfrac{\partial R}{\partial t} = -k_{\mathrm{on}} R\bar{L} + k_{\mathrm{off}} C - k_{\mathrm{er}} R + Q_r \\ \dfrac{\partial S}{\partial t} = k_{\mathrm{s}} C_{\mathrm{i}} - k_{\mathrm{d}} S \end{cases} \tag{1.1}$$

where:

- k_{ec} is the ligand-induced internalization rate constant (min^{-1}),
- k_{on} is the receptor-ligand association constant ($M^{-1}min^{-1} = mol^{-1}cm^3min^{-1}$),
- k_{off} is the receptor-ligand dissociation constant (min^{-1}),
- R_0 is the number of receptors per cell surface area in the absence of ligand ($mol\ cm^{-2}$),
- Q_r is the receptor production rate ($M\ min^{-1}$),
- α_{rec} and α_{deg} are respectively the fraction of recycled and degraded receptors,

- k_{rec} and k_{deg} are respectively the receptor recycling and degradation rates (min^{-1}),
- k_{d} is the degradation rate of dpERK,
- k_{s} is the production rate of dpERK.

Diffusion

Following the work of Goentoro et al [45], we approximate the thin three-dimensional pervitelline space by a 2-dimensional time-varying prolate spheroid (Fig.1.3a). Its dimensions at each stage are extrapolated from experimental measurements (see Table 1.2 and Figure 1.5a).

To derive the reaction-diffusion equation describing the evolution of the concentration of Gurken on the prolate spheroidal surface, we begin by considering the perivitelline space as a three dimensional space enclosed between two surfaces that we denote by $\mathcal{S}_0$ (at the oocyte boundary) and $\mathcal{S}_{\text{FC}}$ (at the follicle cells boundary).

Definition 1.1.1. *Let $a > 0$ and $\xi_0 > 0$. We denote by Φ_0 the map:*

$$\Phi_0 : (\eta, \theta) \in [0, \pi] \times [0, 2\pi] \mapsto \begin{pmatrix} a \sinh \xi_0 \sin \eta \cos \theta \\ a \sinh \xi_0 \sin \eta \sin \theta \\ a \cosh \xi_0 \cos \eta \end{pmatrix} \in \mathbb{R}^3.$$

We denote by $\mathcal{S}_0$ the prolate spheroid parametrized by Φ_0, i.e. $\mathcal{S}_0 := \{\Phi_0(\eta, \theta) \mid (\eta, \theta) \in [0, \pi] \times [0, 2\pi]\}$.

Now the surface $\mathcal{S}_{\text{FC}}$ is assumed to be at a constant distance H from the surface $\mathcal{S}_0$. We define a new map to parametrize each point of the perivitelline space enclosed by $\mathcal{S}_{\text{FC}}$ and $\mathcal{S}_0$:

Definition 1.1.2. *Let $a > 0$ and $\xi_0 > 0$. Let $n : [0, \pi] \times [0, 2\pi] \to \mathbb{S}^2$ denote the function that maps each point (η, θ) of $\mathcal{S}_0$ to the unit normal vector to $\mathcal{S}_0$ at that point, i.e. $n(\eta, \theta) = \frac{\partial_\eta \Phi_0(\eta,\theta) \wedge \partial_\theta \Phi_0(\eta,\theta)}{\|\partial_\eta \Phi_0(\eta,\theta) \wedge \partial_\theta \Phi_0(\eta,\theta)\|}$. We denote by $\Phi^\epsilon : [0, \pi] \times [0, 2\pi] \times [0, H] \to \mathbb{R}^3$ the map defined by:*

$$\Phi^\epsilon(\eta, \theta, \epsilon) = \Phi_0(\eta, \theta) + \epsilon\, n(\eta, \theta).$$

We denote by $\mathcal{S}_{FC}$ the manifold defined by: $\mathcal{S}_{FC} := \{\Phi^\epsilon(\eta, \theta, H) \mid (\eta, \theta) \in [0, \pi] \times [0, 2\pi]\}$.

Remark 1.1.1. *Notice that manifold $\mathcal{S}_0$ can also be defined using the map Φ^ϵ, with: $\mathcal{S}_0 = \{\Phi^\epsilon(\eta, \theta, 0) \mid (\eta, \theta) \in [0, \pi] \times [0, 2\pi]\}$. However, the two manifolds are of different geometric natures. Indeed, contrarily to $\mathcal{S}_0$, $\mathcal{S}_{FC}$ is not a prolate spheroid.*

Remark 1.1.2. *The parameters a and ξ_0 are calculated from the measured dimensions of the oocyte, with $L_{AP} = a\cosh\xi_0$ and $L_{DV} = a\sinh\xi_0$, see Figure 1.3a and Section 1.1.5.*

One can verify that the metric tensor associated with these coordinates is a diagonal matrix of the form:

$$G^\epsilon(\eta,\theta,\epsilon) = (g^\epsilon_{ij})_{1\le i,j\le 3} = \begin{pmatrix} {h^\epsilon_\eta}^2 & 0 & 0 \\ 0 & {h^\epsilon_\theta}^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

where h^ϵ_η and h^ϵ_θ are the scaling factors with respect to θ and η. Notice that the scaling factor with respect to ϵ is equal to 1.

As in [101, 102] we model the diffusion of the ligand L in the three-dimensional perivitelline space (PVS) as follows:

$$\frac{\partial L}{\partial t} = D\Delta L \tag{1.2}$$

where D denotes the diffusion rate, and $\Delta\cdot$ denotes the Laplace operator in $\mathbb{R}^3$.

Equation (1.2) is supplemented with boundary conditions for L on the PVS/follicle cells and PVS/oocyte surfaces:

$$\begin{cases} \left(D\frac{\partial L}{\partial \epsilon} - k_{\text{on}}RL\right)|_{\epsilon=H} = -k_{\text{off}}C \quad \text{at the follicle cells boundary } \epsilon = H \\ D\frac{\partial L}{\partial \epsilon}|_{\epsilon=0} = qV \quad \text{at the oocyte boundary } \epsilon = 0 \end{cases} \tag{1.3}$$

where q is the source function, equal to 1 at the source location and 0 elsewhere (dimensionless), and V is the flux of ligand ($mol\ cm^{-2}min^{-1}$).

As done in [45] and [101], we consider that the perivitelline space height H is negligible compared to the other dimensions of the problem, see Figures 1.1 and 1.3b. Thus, in the following we make the approximation that the ligand diffuses on the 2-dimensional surface of the oocyte. We now introduce the Laplace-Beltrami operator.

Definition 1.1.3. *The Laplace-Beltrami operator is a generalization of the Laplacian for general Riemannian manifolds. Like the Laplacian, it is defined as the divergence of the gradient:*

$$\Delta_{LB} f = \nabla\cdot\nabla f.$$

Let M be a m-dimensional manifold, embedded in $\mathbb{R}^n$, with metric tensor g. Then the Laplace-

Beltrami can be expressed in local coordinates as:

$$\Delta_{LB} f = \frac{1}{\sqrt{|g|}} \sum_{i=1}^{m} \partial_i (\sum_{j=1}^{m} \sqrt{|g|} g^{ij} \partial_j f)$$

where $|g|$ denotes the determinant of the metric tensor and g^{ij} denotes the i, j-th component of the inverse of g.

We also recall the definition of the metric tensor of a Riemannian manifold:

Definition 1.1.4. *Let M be an m-dimensional Riemannian manifold embedded in $\mathbb{R}^n$, equipped with the metric g, denoted at each point $p \in M$ by: $g_p : T_pM \times T_pM \to \mathbb{R}$. Let M be given by the smooth parametrization: $\phi : (x_1, \ldots, x_m) \in \mathcal{U} \subset \mathbb{R}^m \mapsto (\phi_1, \ldots, \phi_n) \in \mathbb{R}^n$, where $\mathcal{U}$ is an open subset of $\mathbb{R}^m$. Let $J\phi$ denote the Jacobian matrix of ϕ. Then the metric tensor at point $p \in M$ is defined by: $G_p := (J\phi_{\phi^{-1}(p)})^T (J\phi_{\phi^{-1}(p)})$.*

We now compute the Laplace-Beltrami operator on $\mathcal{S}_0$.

Lemma 1.1.1. *The Laplace-Beltrami operator of $\mathcal{S}_0$ expressed in the prolate spheroidal coordinates given in Def. 1.1.1 is:*

$$\begin{aligned}\Delta_{LB} f =& \frac{1}{a^2(\sinh^2 \xi_0 + \sin^2 \eta)} \partial_\eta^2 f + \frac{1}{a^2 \sqrt{\sinh^2 \xi_0 + \sin^2 \eta} \sinh \xi_0 \sin \eta} (\partial_\eta \frac{\sinh^2 \xi_0 \sin^2 \eta}{\sqrt{\sinh^2 \xi_0 + \sin^2 \eta}}) \partial_\eta f \\ &+ \frac{1}{a^2 \sinh^2 \xi_0 \sin^2 \eta} \partial_\theta^2 f.\end{aligned} \tag{1.4}$$

Proof. We study the manifold $\mathcal{S}_0$ parametrized by Φ_0 defined in Def. 1.1.1. Because the coordinate vectors ∂_η and ∂_ϕ are orthogonal, we have: $\partial_\eta \Phi_0 \cdot \partial_\theta \Phi_0 = 0$ (see Def. 1.1.4). So G_p simplifies to:

$$G_p = \begin{pmatrix} (h_\eta^0)^2 & 0 \\ 0 & (h_\theta^0)^2 \end{pmatrix}, \tag{1.5}$$

where h_η^0 and h_θ^0 are the scaling factors, defined by: $h_\eta^0 = \sqrt{\partial_\eta \Phi_0 \cdot \partial_\eta \Phi_0}$ and $h_\theta^0 = \sqrt{\partial_\theta \Phi_0 \cdot \partial_\theta \Phi_0}$.

The diagonal nature of the metric tensor allows us to compute explicitly:

$$\begin{aligned}\Delta_{\mathrm{LB}} f &= \frac{1}{\sqrt{|g|}} \sum_{i=1}^{m} \partial_i (\sum_{j=1}^{m} \sqrt{|g|} g^{ij} \partial_j f) \\ &= \frac{1}{h_\eta^0 h_\theta^0} \left(\partial_\eta (h_\eta^0 h_\theta^0 (h_\eta^0)^{-2} \partial_\eta f) + \partial_\theta (h_\eta^0 h_\theta^0 (h_\theta^0)^{-2} \partial_\theta f) \right) \\ &= \frac{1}{h_\eta^0 h_\theta^0} \left(\partial_\eta (\frac{h_\theta^0}{h_\eta^0} \partial_\eta f) + \partial_\theta (\frac{h_\eta^0}{h_\theta^0} \partial_\theta f) \right).\end{aligned} \tag{1.6}$$

The scaling factors can be calculated:

$$\begin{cases} h_\eta^0 = \sqrt{\partial_\eta \Phi_0 \cdot \partial_\eta \Phi_0} = \sqrt{a^2 \sinh^2 \xi_0 \cos^2 \eta + a^2 \cosh^2 \xi_0 \sin^2 \eta} = a\sqrt{\sinh^2 \xi_0 + \sin^2 \eta} \\ h_\theta^0 = \sqrt{\partial_\theta \Phi_0 \cdot \partial_\theta \Phi_0} = a \sinh \xi_0 \sin \eta \end{cases}$$

Notice that the scaling factors do not depend on θ. Hence the Laplace-Beltrami operator rewrites as:

$$\Delta_{\mathrm{LB}} f = \frac{1}{h_\eta^0 h_\theta^0} \partial_\eta (\frac{h_\theta^0}{h_\eta^0} \partial_\eta f) + \frac{1}{(h_\theta^0)^2} \partial_\theta^2 f = \frac{1}{(h_\eta^0)^2} \partial_\eta^2 f + \frac{1}{h_\eta^0 h_\theta^0} (\partial_\eta \frac{h_\theta^0}{h_\eta^0}) \partial_\eta f + \frac{1}{(h_\theta^0)^2} \partial_\theta^2 f. \tag{1.7}$$

□

Theorem 1.1.1. *Let $L \in C^2([0,\pi] \times [0,2\pi] \times [0,H])$, satisfying Equations (1.2) and (1.3). Suppose that the perivitelline space height H is small enough that L, h_η^ϵ and h_θ^ϵ do not very appreciably along the coordinate ϵ, i.e. for all $(\eta,\theta,\epsilon) \in [0,\pi] \times [0,2\pi] \times [0,H]$, $L(\eta,\theta,\epsilon) = L(\eta,\theta,0)$, $h_\eta^\epsilon = h_\eta^0$ and $h_\theta^\epsilon = h_\theta^0$. Let $\tilde{L} := \int_0^H L d\epsilon$. Then integrating equation (1.2) between $\epsilon = 0$ and $\epsilon = H$ yields:*

$$\frac{\partial \tilde{L}}{\partial t} = D \Delta_{LB} \tilde{L} - \frac{1}{H} k_{on} R \tilde{L} + k_{off} C + qV. \tag{1.8}$$

Proof. Denoting $|g^\epsilon| = \det(g^\epsilon)$, the diffusion operator can be written as:

$$\Delta L = \frac{1}{\sqrt{|g^\epsilon|}} \frac{\partial}{\partial \epsilon} \left(\sqrt{|g^\epsilon|} \frac{\partial L}{\partial \epsilon} \right) + \Delta_{\mathrm{surf}} L$$

where $\Delta_{\mathrm{surf}} L$ denotes the term involving the surface derivatives $\frac{\partial}{\partial \eta}$ and $\frac{\partial}{\partial \theta}$, i.e.

$$\Delta_{\mathrm{surf}} L = \frac{1}{\sqrt{|g^\epsilon|}} \left[\frac{\partial}{\partial \eta} \left(\frac{\sqrt{|g^\epsilon|}}{(h_\eta^\epsilon)^2} \frac{\partial L}{\partial \eta} \right) + \frac{\partial}{\partial \theta} \left(\frac{\sqrt{|g^\epsilon|}}{(h_\theta^\epsilon)^2} \frac{\partial L}{\partial \theta} \right) \right]$$

Due to the particular form of the metric tensor, this rewrites as:

$$\Delta_{\text{surf}} L = \frac{1}{h_\eta^\epsilon h_\theta^\epsilon}\left[\frac{\partial}{\partial\eta}\left(\frac{h_\eta^\epsilon h_\theta^\epsilon}{(h_\eta^\epsilon)^2}\frac{\partial L}{\partial\eta}\right) + \frac{\partial}{\partial\theta}\left(\frac{h_\eta^\epsilon h_\theta^\epsilon}{(h_\theta^\epsilon)^2}\frac{\partial L}{\partial\theta}\right)\right] = \frac{1}{h_\eta^\epsilon h_\theta^\epsilon}\left[\frac{\partial}{\partial\eta}\left(\frac{h_\theta^\epsilon}{h_\eta^\epsilon}\frac{\partial L}{\partial\eta}\right) + \frac{\partial}{\partial\theta}\left(\frac{h_\eta^\epsilon}{h_\theta^\epsilon}\frac{\partial L}{\partial\theta}\right)\right].$$

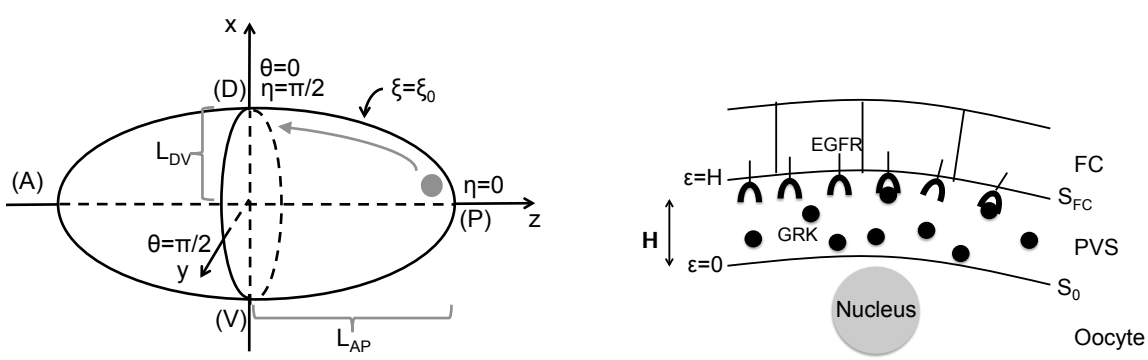
Since $h_\eta^\epsilon = h_\eta^0$ and $h_\theta^\epsilon = h_\theta^0$, we recognize the Laplace-Beltrami operator of $\mathcal{S}_0$, i.e. $\Delta_{\text{surf}} = \Delta_{\text{LB}}$. Integrating Equation (1.2) between $\epsilon = 0$ and $\epsilon = H$ then yields:

$$\frac{\partial L}{\partial t} = D\Delta_{\text{LB}} L + \frac{1}{H}(-k_{\text{on}} RL + k_{\text{off}} C + qV). \tag{1.9}$$

Let $\tilde{L} := \int_0^H L d\epsilon = LH$. This yields (1.8). □

Remark 1.1.3. *In [45], the surface separating the perivitelline space from the follicle cells is parametrized by a slightly larger prolate spheroid. Consequently, when averaging L between the two surfaces, the authors obtain an operator that differs from the Laplace-Beltrami one. This operator does not conserve mass, whereas the Laplace-Beltrami operator does. For this reason, we use Equation* (1.8) *instead of the equation given in [45].*

For simplicity of notation, we will now denote by L the surface concentration of ligand introduced as $\tilde{L}$.



(a) Prolate spheroidal coordinates for the *Drosophila* oocyte

(b) Coordinates at the boundaries of the perivitelline space

Figure 1.3: Oocyte as a prolate spheroid

Negative feedback

In addition to the mechanisms described in sections 1.1.1 and 1.1.1, we consider several feedback loops.

Recycling of receptors As represented in Equations (1.1) and (1.20), a fraction α_{deg} of the internalized receptors R_{i} goes to degradation, while the fraction $\alpha_{\text{rec}} = 1 - \alpha_{\text{deg}}$ is recycled and goes back to the membranes of the follicle cells to be reused. In [115], Sigismund et al. showed that there exist two different pathways for the EGFR internalization: clathrin-regulated endocytosis (CME) and non-clathrin-mediated endocytosis (NCE).

While 70% of the receptors internalized through the CME pathway are recycled, only 15% of the receptors undergoing NCE are recycled [115]. Importantly, at low level of ligand, almost all receptors undergo clathrin-mediated endocytosis. At high level of ligand, 60% of EGFR undergo CME and 40% undergo NCE. We incorporate this data in our model, considering that the fractions of degraded and recycled receptors depend on the level of ligand. At low ligand, $\alpha_{\text{deg}} = 0.3$, and at high ligand, $\alpha_{\text{deg}} = 0.55$. At intermediate level of ligand, we interpolate linearly as follows:

$$\alpha_{\text{deg}}(t,\eta,\theta) = 0.3 + 0.25\frac{L(t,\eta,\theta) - L_{\min}(t)}{L_{\max}(t) - L_{\min}(t)}, \quad \alpha_{\text{rec}}(t,\eta,\theta) = 1 - \alpha_{\text{deg}}(t,\eta,\theta) \tag{1.10}$$

where $L_{\min}(t)$ and $L_{\max}(t)$ are respectively the minimum and maximum values of L at time t:

$$L_{\min}(t) = \min\{L(t,\eta,\theta) \mid (\eta,\theta) \in [0,\pi] \times [0,2\pi]\},\ L_{\max}(t) = \max\{L(t,\eta,\theta) \mid (\eta,\theta) \in [0,\pi] \times [0,2\pi]\}.$$

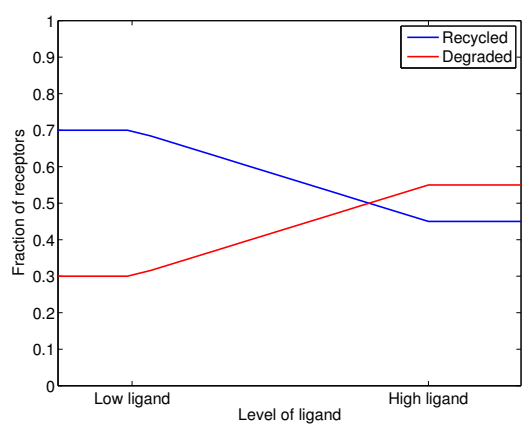


Figure 1.4: Fractions of recycled and degraded receptors as functions of the level of ligand.

Action of inhibitors We consider the action of two inhibitors of the EGFR pathway, Kekkon1 (Kek1) and Sprouty (Sty), see Figure 1.1. While both inhibitors share similar spatial domains of expression, the repressive mechanism of each inhibitor is different [92]. The transmembrane protein

Kek1 directly interacts with the EGF receptor to inhibit ligand-receptor interactions [42]. On the other hand, Sty acts directly on Ras/MAPK to inhibit dpERK activation [92, 114]. Each inhibitor has to be considered independently based on its inhibitory mechanism.

Kek1 targets EGFR dimerization, thus reducing GRK uptake and leaving higher levels of free ligand. To account for this effect, we modify the receptor-ligand binding rate via Michaelis-Menten kinetics. Let I_{K} denote the space and time-varying concentration of Kek1. Then we write:

$$\tilde{k}_{\mathrm{on}}(t,\eta,\theta) = \frac{k_{\mathrm{on}}}{1+\gamma_{\mathrm{Kek}} I_{\mathrm{K}}(t,\eta,\theta)/\bar{S}} \tag{1.11}$$

The parameter γ_{Kek} is the strength of Kek1's inhibitory feedback, and the constant $\bar{S}$ is defined by: $\bar{S} = \frac{V k_{\mathrm{s}}}{k_{\mathrm{d}} k_{\mathrm{deg}}}$. The new binding rate $\tilde{k}_{\mathrm{on}}$ is now a space and time dependent variable, affected by Kek1. In the absence of Kek1, $\tilde{k}_{\mathrm{on}} \equiv k_{\mathrm{on}}$ and we recover the constant binding rate previously defined (See Table 1.1).

Sty acts on the intracellular components, affecting signal propagation (Fig. 1.1). We model its effect by modifying the internalization rate of dpERK, also via Michaelis-Menten kinetics. Let I_{S} denote the concentration of Sty. We define:

$$\tilde{k}_{\mathrm{s}}(t,\eta,\theta) = \frac{k_{\mathrm{s}}}{1+\gamma_{\mathrm{Sty}} I_{\mathrm{S}}(t,\eta,\theta)/\bar{S}} \tag{1.12}$$

where γ_{Sty} is the strength of inhibitory feedback.

In turn, as targets of the pathway, the concentrations of Sty and Kek1 depend on the concentration of dpERK and are modeled by standard linear kinetics:

$$\begin{cases} \dfrac{\partial I_{\mathrm{K}}}{\partial t} = k^{\mathrm{Kek}} S - k_d^{\mathrm{Kek}} I_{\mathrm{K}} \\ \dfrac{\partial I_{\mathrm{S}}}{\partial t} = k^{\mathrm{Sty}} S - k_d^{\mathrm{Sty}} I_{\mathrm{S}} \end{cases} \tag{1.13}$$

where k^{Sty}, k_d^{Sty}, k^{Kek} and k_d^{Kek} are the production rates and degradation rates of Sty and Kek1.

With this added mechanism, the complete dynamics form the following system of coupled PDE-

ODEs:

$$\begin{cases} \dfrac{\partial L}{\partial t} = D\Delta_{\mathrm{LB}} L - \frac{1}{H}\tilde{k}_{\mathrm{on}} RL + k_{\mathrm{off}} C + qV \\ \dfrac{\partial C}{\partial t} = \frac{1}{H}\tilde{k}_{\mathrm{on}} RL - (k_{\mathrm{off}} + k_{\mathrm{ec}})C + \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} \\ \dfrac{\partial C_{\mathrm{i}}}{\partial t} = k_{\mathrm{ec}} C - \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} - \alpha_{\mathrm{deg}} k_{\mathrm{deg}} C_{\mathrm{i}} \\ \dfrac{\partial R}{\partial t} = -\frac{1}{H}\tilde{k}_{\mathrm{on}} RL + k_{\mathrm{off}} C - k_{\mathrm{er}} R + Q_r \\ \dfrac{\partial S}{\partial t} = \tilde{k}_{\mathrm{s}} C_{\mathrm{i}} - k_{\mathrm{d}} S \end{cases} \tag{1.14}$$

1.1.2 Growth of the egg chamber

In the model described by equations (1.14)-(1.11)-(1.12)-(1.13), the growth of the oocyte is not taken into account. However, experimental measurements show that as oogenesis progresses from Stage 7 to Stage 10A, the anterior-posterior dimensions of the egg-chamber increase by a factor of 4 and the dorso-ventral ones by a factor of 3 (see Table 1.2). This hints that growth may play a fundamental role in shaping the distributions of ligand and signal. Using our new framework of Developmental PDEs (see Chapter 2), we now include the evolution of the shape of the domain.

Let $v \in \mathrm{Lip}(\mathbb{R}^3, \mathbb{R}^3)$ be a Lipshitz vector field. The perivitelline space, approximated by a 2-dimensional compact manifold embedded in $\mathbb{R}^3$ now considered to vary with time, is denoted by $\mathcal{S}_t$. Its evolution can be described as the push-forward of the initial manifold via the vector field v. Let us denote by ϕ_v^t the flow of v at time t. Then, denoting by $\mathcal{S}_0$ the manifold at time 0, $\mathcal{S}_t$ is given by:

$$\mathcal{S}_t = \phi_v^t \# \mathcal{S}_0. \tag{1.15}$$

We rewrite equations (1.14) and (1.13) as follows:

$$\begin{cases} \dfrac{\partial L}{\partial t} = D\Delta_{\mathrm{LB}} L + \nabla \cdot (vL) - \frac{1}{H}\tilde{k}_{\mathrm{on}} RL + k_{\mathrm{off}} C + qV \\ \dfrac{\partial C}{\partial t} = \nabla \cdot (vC) + \frac{1}{H}\tilde{k}_{\mathrm{on}} RL - (k_{\mathrm{off}} + k_{\mathrm{ec}})C + \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} \\ \dfrac{\partial C_{\mathrm{i}}}{\partial t} = \nabla \cdot (vC_{\mathrm{i}}) + k_{\mathrm{ec}} C - \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} - \alpha_{\mathrm{deg}} k_{\mathrm{deg}} C_{\mathrm{i}} \\ \dfrac{\partial R}{\partial t} = \nabla \cdot (vR) - \frac{1}{H}\tilde{k}_{\mathrm{on}} R\bar{L} + k_{\mathrm{off}} C - k_{\mathrm{er}} R + Q_r \\ \dfrac{\partial S}{\partial t} = \nabla \cdot (vS) + \tilde{k}_{\mathrm{s}} C_{\mathrm{i}} - k_{\mathrm{d}} S \end{cases} \tag{1.16}$$

and

$$\begin{cases} \dfrac{\partial I_{\mathrm{K}}}{\partial t} = \nabla \cdot (v I_{\mathrm{K}}) + k^{\mathrm{Kek}} S - k_d^{\mathrm{Kek}} I_{\mathrm{K}} \\ \dfrac{\partial I_{\mathrm{S}}}{\partial t} = \nabla \cdot (v I_{\mathrm{S}}) + k^{\mathrm{Sty}} S - k_d^{\mathrm{Sty}} I_{\mathrm{S}}. \end{cases} \tag{1.17}$$

Notice that this is a modified version of the equation developed in Chapter 2. Indeed, the vector field v does not depend on the measure diffusing on its surface. Furthermore, the quantities L, R, C, C_{i}, S, I_{K} and I_{S} have time-varying mass, hence they cannot be considered as probability measures. To give a rigorous mathematical definition of equations (1.16) and (1.17), it is necessary to extend the framework of DPDEs. This is a future direction of this thesis.

1.1.3 Shift of the follicle cells

The follicle cells overlaying the perivitelline space are known to gradually shift from the anterior to the posterior of the egg chamber (see Figure 1.1) [118]. Since the receptors, complexes, signal and inhibitors are located inside or on the membrane of the follicle cells, they are affected by this movement. This phenomenon can be transcribed mathematically by adding a transport term to the equations of these variables. We introduce a time-dependent vector field tangent to the surface of the prolate spheroid. Let $w_t \in \mathrm{Lip}(\mathcal{S}_t, T\mathcal{S}_t)$. The full set of equations including the phenomena of growth and shift of the follicle cells rewrites:

$$\begin{cases} \dfrac{\partial L}{\partial t} = D\Delta_{\mathrm{LB}} L + \nabla \cdot (vL) - \frac{1}{H}\tilde{k}_{\mathrm{on}} RL + k_{\mathrm{off}} C + qV \\ \dfrac{\partial C}{\partial t} = \nabla \cdot (vC) + \nabla \cdot (w_t C) + \frac{1}{H}\tilde{k}_{\mathrm{on}} RL - (k_{\mathrm{off}} + k_{\mathrm{ec}})C + \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} \\ \dfrac{\partial C_{\mathrm{i}}}{\partial t} = \nabla \cdot (vC_{\mathrm{i}}) + \nabla \cdot (w_t C_{\mathrm{i}}) + k_{\mathrm{ec}} C - \alpha_{\mathrm{rec}} k_{\mathrm{rec}} C_{\mathrm{i}} - \alpha_{\mathrm{deg}} + k_{\mathrm{deg}} C_{\mathrm{i}} \\ \dfrac{\partial R}{\partial t} = \nabla \cdot (vR) + \nabla \cdot (w_t R) - \frac{1}{H}\tilde{k}_{\mathrm{on}} R\bar{L} + k_{\mathrm{off}} C - k_{\mathrm{er}} R + Q_r \\ \dfrac{\partial S}{\partial t} = \nabla \cdot (vS) + \nabla \cdot (w_t S) + \tilde{k}_{\mathrm{s}} C_{\mathrm{i}} - k_{\mathrm{d}} S \end{cases} \tag{1.18}$$

and

$$\begin{cases} \dfrac{\partial I_{\mathrm{K}}}{\partial t} = \nabla \cdot (v I_{\mathrm{K}}) + \nabla \cdot (w_t I_{\mathrm{K}}) + k^{\mathrm{Kek}} S - k_d^{\mathrm{Kek}} I_{\mathrm{K}} \\ \dfrac{\partial I_{\mathrm{S}}}{\partial t} = \nabla \cdot (v I_{\mathrm{S}}) + \nabla \cdot (w_t I_{\mathrm{S}}) + k^{\mathrm{Sty}} S - k_d^{\mathrm{Sty}} I_{\mathrm{S}}. \end{cases} \tag{1.19}$$

1.1.4 Rescaling

We rescale Equations (1.18)-(1.19) by the quantities $L_0 = HV/k_{\mathrm{I}}$, $C_0 = V/k_{\mathrm{ec}}$, R_0 and $S_0 = V/k_{\mathrm{d}}$. These constants respectively represent the concentrations of ligand, complexes, receptors and signal in the absence of spatial and temporal variation (i.e. setting all spatial and temporal derivatives to 0). The constant k_I defined by $k_I = (k_{\mathrm{ec}}k_{\mathrm{on}}R)/(k_{\mathrm{off}} + k_{\mathrm{ec}})$ is the rate of internalization of ligand at steady-state. Its fundamental role in determining the shape of the signal at steady-state was discussed in [45].

Rescaling by L_0, C_0, R_0 and S_0 renders the variables dimensionless. It also ensures that the new variables $l = L/L_0$, $c = C/C_0$, $c_{\mathrm{i}} = C_{\mathrm{i}}/C_0$, $r = R/R_0$, $s = S/S_0$, $i_{\mathrm{K}} = I_{\mathrm{K}}/S_0$ and $i_{\mathrm{S}} = I_{\mathrm{S}}/S_0$ are of the order of 1, which allows greater numerical precision. We rewrite the system of equations (1.14) in terms of the dimensionless distributions l, c, c_{i}, r and s:

$$\begin{cases} \dfrac{\partial l}{\partial t} = D\Delta_{\mathrm{LB}} l + \nabla\cdot(vl) - \frac{1}{H}\tilde{k}_{\mathrm{on}} R_0 r l + k_{\mathrm{off}}\frac{C_0}{L_0} c + q\frac{V}{L_0} \\ \dfrac{\partial c}{\partial t} = \nabla\cdot(vc) + \nabla\cdot(w_t c) + \tilde{k}_{\mathrm{on}}\frac{R_0 L_0}{H C_0} r l - (k_{\mathrm{off}} + k_{\mathrm{ec}})c + \alpha_{\mathrm{rec}} k_{\mathrm{rec}} c_{\mathrm{i}} \\ \dfrac{\partial c_{\mathrm{i}}}{\partial t} = \nabla\cdot(vc_{\mathrm{i}}) + \nabla\cdot(w_t c_{\mathrm{i}}) + k_{\mathrm{ec}} c - \alpha_{\mathrm{rec}} k_{\mathrm{rec}} c_{\mathrm{i}} - \alpha_{\mathrm{deg}} k_{\mathrm{deg}} c_{\mathrm{i}} \\ \dfrac{\partial r}{\partial t} = \nabla\cdot(vr) + \nabla\cdot(w_t r) - \frac{1}{H}\tilde{k}_{\mathrm{on}} L_0 r l + k_{\mathrm{off}}\frac{C_0}{R_0} c - k_{\mathrm{er}} r + \frac{Q_r}{R_0} \\ \dfrac{\partial s}{\partial t} = \nabla\cdot(vs) + \nabla\cdot(w_t s) + \tilde{k}_{\mathrm{s}}\frac{C_0}{S_0} c_{\mathrm{i}} - k_{\mathrm{d}} s. \end{cases} \tag{1.20}$$

and

$$\begin{cases} \dfrac{\partial i_{\mathrm{K}}}{\partial t} = \nabla\cdot(vi_{\mathrm{K}}) + \nabla\cdot(w_t i_{\mathrm{K}}) + k^{\mathrm{Kek}} s - k_d^{\mathrm{Kek}} i_{\mathrm{K}} \\ \dfrac{\partial i_{\mathrm{S}}}{\partial t} = \nabla\cdot(vi_{\mathrm{S}}) + \nabla\cdot(w_t i_{\mathrm{S}}) + k^{\mathrm{Sty}} s - k_d^{\mathrm{Sty}} i_{\mathrm{S}}. \end{cases} \tag{1.21}$$

1.1.5 Calibration of the model

The model parameters are carefully chosen taking values from the literature. The following table summarizes the values that we use and their justification:

Source of ligand Gurken RNA is secreted from the nurse cells on the anterior of the oocyte and it gets localized around the oocyte nucleus. The source of ligand can be approximated from images of Gurken RNA in *Drosophila melanogaster*. We model it as a triangular shape with dimensions

Parameter	Definition	Typical Value or Range	Reference
H	Perivitelline space thickness	$0.5\ \mu m$	Measurements
D	Diffusion rate	$3,600 - 360,000\ \mu m^2\ hr^{-1}$ $36 - 360,000\ \mu m^2\ hr^{-1}$	[102] [22]
k_{ec}	Complex internalization rate	$6\ hr^{-1}$	[102]
k_{on}	Receptor-ligand association rate	$10^{22} - 10^{25}\ mol^{-1}\mu m^3 hr^{-1}$	[102]
k_{off}	Receptor-ligand dissociation rate	$6\ hr^{-1}$	[102]
R_0	Number of receptors per surface area in the absence of ligand	$6.7 \times 10^{22} mol\ \ \mu m^{-2}$	[102]
k_{er}	Free receptor internalization rate	$0.6 - 6\ hr^{-1}$	[102]
α_{rec}	Fraction of recycled receptors	$0.45 - 0.7$	[115]
α_{deg}	Fraction of degraded receptors	$0.3 - 0.55$	[115]
k_{rec}	Receptor recycling rate	$2.3\ hr^{-1}$	[115]
k_{deg}	Receptor degradation rate	$2.3\ hr^{-1}$	[115]
k_{d}	dpERK degradation rate	$2.5\ hr^{-1}$	[97]

Table 1.1: Justification of the chosen values of parameters.

Stage	S7	S8(E)	S8(L)	S9(E)	S9(L)	S10A
Time (hr)	3	7.5	10.5	13.5	16.5	19.5
L_{AP} (μm)	71	99	132	190	246	304
L_0 (μm)	-	19	31	63	111	152
L_{FC} (μm)	71	99	132	127	131	152
L_{source}/L_0	1	-	0.4	0.4	0.4	0.4

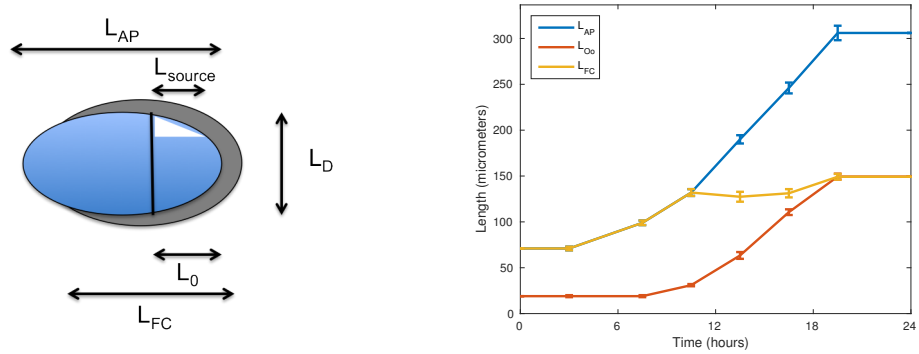
Table 1.2: Measured dimensions of the egg chamber. See Figure 1.5a for schematic of measurements.

based on experimental measurements at each stage (see Table 1.2 and Figure 1.5a).

Time-varying dimensions The evolving dimensions of the egg chamber such as anterior-posterior length, dorso-ventral length, length of follicle cells, length of the oocyte and dimensions of the source were measured at different stages of oogenesis (see Figure 1.5a and Table 1.2). Then they were interpolated so as to get a continuous description of the dimensions over time (see Figure 1.5b). These measurements allowed us to calibrate the growth vector field v and the cell shift vector field w_t.

1.2 Numerics

The complete model that we are studying is composed of several components that each pose numerical challenges in specific ways. Solving Equation (1.20) numerically requires finding a suitable spatial discretization of the domain, in this case a two-dimensional prolate spheroid. The most natural parametrization of such a symmetric surface is done with the prolate spheroidal coordinates. However, this system of coordinates is degenerate at the two poles of the spheroid. As a conse-



(a) Schematic of the measurements of the egg chamber. Measurements were done at each stage (see Table 1.2).

(b) Interpolation of the measurements.

Figure 1.5: Measurements of the egg chamber's dimensions

quence, the corresponding mesh constructed with prolate spheroidal coordinates is ill-suited for the numerical approximation of diffusion. As an alternative to prolate spheroidal coordinates, we used cubed spheroidal coordinates, adapted from the cubed sphere coordinates developped in [47, 83]. Before exploring in detail the numerics of each component of the model, we first describe these two possible spatial discretizations.

1.2.1 Spheroidal and cubed spheroidal parametrizations

Let $(\eta, \theta) \in [0, \pi] \times [0, 2\pi]$. Let $(L_{\mathrm{DV}}, L_{\mathrm{AP}}) \in (\mathbb{R}+)^2$ denote the half lengths of the small and big axes respectively. Let $(a_0, \xi_0) \in (\mathbb{R}^+)^2$ such that

$$L_{\mathrm{DV}} = a_0 \sinh \xi_0 \text{ and } L_{\mathrm{AP}} = a_0 \cosh \xi_0,$$

i.e.

$$\tanh \xi_0 = \frac{L_{\mathrm{DV}}}{L_{\mathrm{AP}}} \text{ and } a_0 = \sqrt{L_{\mathrm{AP}}^2 - L_{\mathrm{DV}}^2}.$$

In what follows, we use both the parameters $(L_{\mathrm{DV}}, L_{\mathrm{AP}})$, coming from the notations of the biological application, and the parameters (a_0, ξ_0).

A prolate spheroid S obtained by rotating an ellipse around its big axis can be described in $\mathbb{R}^3$

by the parametrization $(\eta, \theta) \mapsto (x, y, z)$ with:

$$\begin{cases} x(\eta,\theta) = L_{\mathrm{DV}} \sin\eta \cos\theta = a_0 \sinh\xi_0 \sin\eta \cos\theta \\ y(\eta,\theta) = L_{\mathrm{DV}} \sin\eta \sin\theta = a_0 \sinh\xi_0 \sin\eta \sin\theta \\ z(\eta,\theta) = L_{\mathrm{AP}} \cos\eta = a_0 \cosh\xi_0 \cos\eta, \end{cases} \tag{1.22}$$

where L_{AP} and L_{DV} denote the half lengths of the big and small axes, respectively. This parametrization is not a diffeomorphism from $[0, \pi] \times [0, 2\pi]$ to S. Indeed, notice that for all $\theta \in [0, 2\pi]$, $(x, y, z)(0, \theta) = (0, 0, L_{\mathrm{AP}})$ and $(x, y, z)(\pi, \theta) = (0, 0, -L_{\mathrm{AP}})$.

Spheroidal parametrization

We construct a prolate spheroidal mesh as follows. Let $(N_\eta, N_\theta) \in \mathbb{N}^2$. We define $\Delta\eta = \pi/N_\eta$ and $\Delta\theta = \pi/N_\theta$. For all $i \in \{1, \ldots, N_\eta\}$, for all $j \in \{1, \ldots, N_\theta\}$, let

$$\eta_i = i\Delta\eta, \quad \text{and} \quad \theta_j = j\Delta\theta$$

and we define

$$\begin{cases} x_{ij} = L_{\mathrm{DV}} \sin\eta_i \cos\theta_j \\ y_{ij} = L_{\mathrm{DV}} \sin\eta_i \sin\theta_j \\ z_{ij} = L_{\mathrm{AP}} \cos\eta_i. \end{cases}$$

Due to the non-diffeomorphic coordinates, this mesh contains two singularities, or overlapping points, at $(0, 0, L_{\mathrm{AP}})$ and $(0, 0, -L_{\mathrm{AP}})$. As a consequence, the discretization points close to the poles are much closer than those towards the equator $z = 0$. This characteristic implies that the mesh has very irregular cell sizes, which makes it ill-suited for the finite-differences approximation of the diffusion operator. For this reason, we present another system of coordinates that provides a more regular discretization of the domain.

Cubed spheroidal parametrization

A way to construct a more regular discretization of the spheroid is to divide it into several subdomains, each endowed with their own coordinate system. We developed the *cubed spheroid* coordinate system by extending the "cubed sphere" approach introduced in [47, 83] in the context of the discontinuous Galerkin numerical scheme.

In order to extend the cubed sphere parametrization, we define a homemorphism between each point of the prolate spheroid S of small axis length L_{DV} and big axis length L_{AP} parametrized by (1.22), and the sphere $\mathbb{S}_{L_{\mathrm{AP}}}$ of radius L_{AP} given by the equation $x^2 + y^2 + z^2 = L_{\mathrm{AP}}^2$. Let $\Phi_s \in C(S, \mathbb{S}_{L_{\mathrm{AP}}})$ be defined by:

$$\Phi_s : (x_s, y_s, z_s) \mapsto (x, y, z) = (x_s \tanh \xi_0, y_s \tanh \xi_0, z) = (\frac{L_{\mathrm{DV}}}{L_{\mathrm{AP}}} x_s, \frac{L_{\mathrm{DV}}}{L_{\mathrm{AP}}} y_s, z). \tag{1.23}$$

The homeomorphism $(\Phi_s)^{-1}$ transforms each point P of the prolate spheroid S to a point P_s of the sphere $\mathbb{S}_{L_{\mathrm{AP}}}$ by projection along the direction $\overrightarrow{P_z P}$, where P_z is has coordinates $(0, 0, z)$ (see Figure 1.6a).

The division of the sphere into "cubed" subdomains was introduced in [83]. We recall it here. Let C_a be the cube of radius $2a$ inscribed in $\mathbb{S}_{L_{\mathrm{AP}}}$, oriented such that the 3D Cartesian axes are orthogonal to its faces (see Figure 1.6). By definition, $a = \frac{1}{\sqrt{3}} L_{\mathrm{AP}}$. We define the mapping $\Phi_c : P_s \in \mathbb{S}_{L_{\mathrm{AP}}} \mapsto P_c \in C_a$ by projection along the direction $\overrightarrow{OP_s}$. Let (x_c, y_c) be local coordinates on each face of C_a. Then the point P_s can be parametrized by the coordinates (x_c, y_c) and the parametrization depends on the face of C_a. Let F_0 be the face of C_a belonging to the plane $z = a$. Let $(x_c, y_c) \in [-a, a] \times [-a, a]$ be the local coordinates on F_0 (see Figure 1.6). Geometrically, we have the following relation between (x_c, y_c) and (x_s, y_s, z_s):

$$\begin{cases} x_c = \frac{a x_s}{z_s} \\ y_c = \frac{a y_s}{z_s}. \end{cases}$$

Then $\Phi_c^0 : F_0 \to \mathbb{S}_{L_{\mathrm{AP}}}$ is defined by:

$$\Phi_c^0 : (x_c, y_c) \mapsto (x_s, y_s, z_s) = \left(\frac{x_c L_{\mathrm{AP}}}{\sqrt{a^2 + x_c^2 + y_c^2}}, \frac{y_c L_{\mathrm{AP}}}{\sqrt{a^2 + x_c^2 + y_c^2}}, \frac{a L_{\mathrm{AP}}}{\sqrt{a^2 + x_c^2 + y_c^2}} \right). \tag{1.24}$$

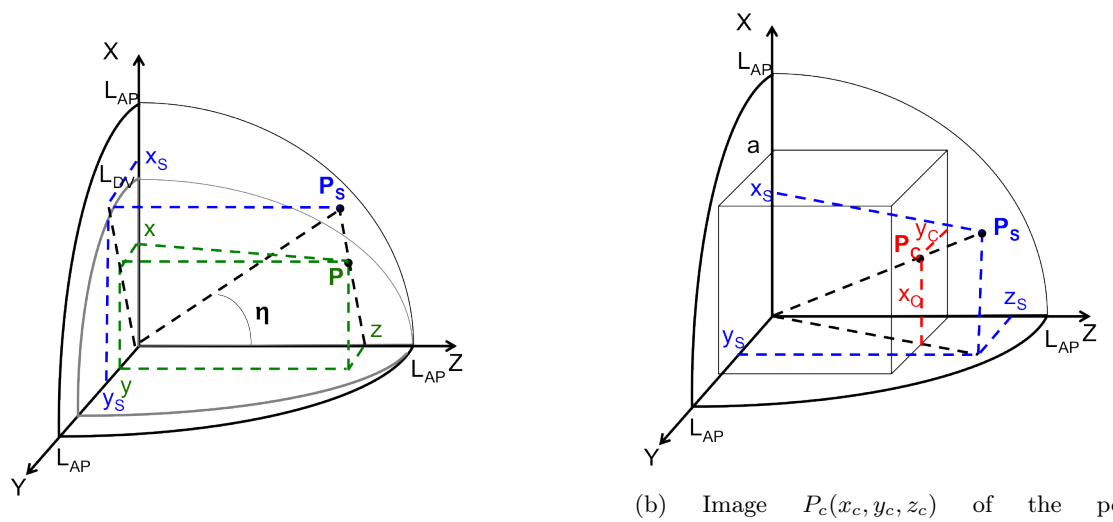
Similar parametrizations can be given by defining local coordinate systems on the other faces of the cube.

Composing Φ_c and Φ_s gives a parametrization of the prolate spheroid S that we name *cubed spheroid parametrization*. The respective images of the faces F_i of C_a by $\Phi_s \circ \Phi_c^i$ (for $i \in \{1, \ldots, 6\}$) divide the spheroid into 6 domains S_i. For instance, the local coordinates on F_0 define a subdomain

of S that we denote by D_0 and that can be parametrized combining (1.23) and (1.24):

$$(x, y, z) = \Phi_s \circ \Phi_c^0(x_c, y_c) = \left(\frac{x_c L_{\mathrm{DV}}}{\sqrt{a^2 + x_c^2 + y_c^2}}, \frac{y_c L_{\mathrm{DV}}}{\sqrt{a^2 + x_c^2 + y_c^2}}, \frac{a L_{\mathrm{AP}}}{\sqrt{a^2 + x_c^2 + y_c^2}}\right). \tag{1.25}$$

In practice, we restrict ourselves to a quarter prolate spheroid (see Figure 1.6a). In this case, only four faces of the cube are needed to parametrize it, which divides the spheroid into four subdomains, denoted by D_0, D_1, D_2 and D_3. The coordinates on the total prolate spheroid can be obtained by symmetry.



(a) Image $P_s(x_s, y_s, z) \in \mathbb{S}_{L_{\mathrm{AP}}}$ of the point $P(x, y, z) \in S$ by the homeomorphism $(\Phi_s)^{-1}$.

(b) Image $P_c(x_c, y_c, z_c)$ of the point $P_s(x_s, y_s, z_s) \in \mathbb{S}_{L_{\mathrm{AP}}}$ by the homeomorphism $(\Phi_c^0)^{-1}$.

Figure 1.6: Construction of the image P_c of $P \in S$ by the homeomorphism $(\Phi_c^0)^{-1} \circ (\Phi_s)^{-1}$.

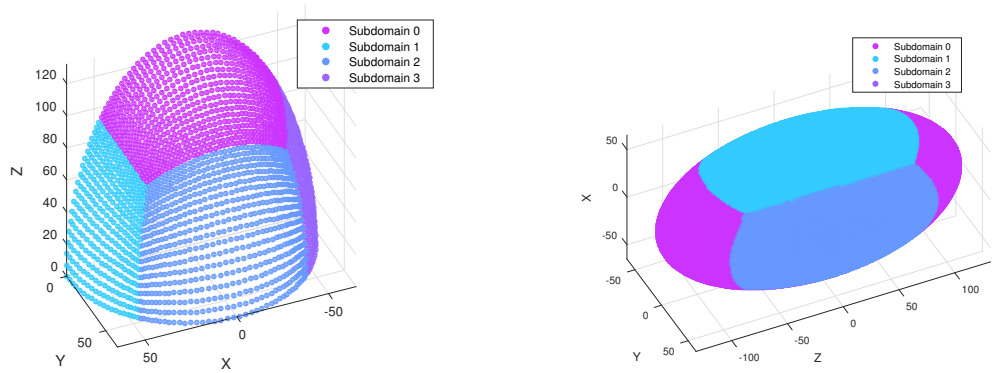
As with the prolate spheroidal parametrization, we create a cubed spheroidal mesh by discretizing the coordinates (x_c, y_c) in each subdomain. As an example, we focus on the subdomain D_0. Let $(N_x^0, N_y^0) \in \mathbb{N}^3$. We define $\Delta x_0 = 2a/N_x$ and $\Delta y_0 = a/N_y$. For all $i \in \{1, \ldots, N_x^0\}$, for all $j \in \{1, \ldots, N_y^0\}$, let

$$x_{ci} = i\Delta x_0, \quad \text{and} \quad y_{cj} = j\Delta y_0$$

and in D_0, we define

$$\begin{cases} x_{ij} = \frac{x_{ci} L_{\mathrm{DV}}}{\sqrt{a^2 + x_{ci}^2 + y_{cj}^2}}, \\ y_{ij} = \frac{y_{cj} L_{\mathrm{DV}}}{\sqrt{a^2 + x_{ci}^2 + y_{cj}^2}}, \\ z_{ij} = \frac{a L_{\mathrm{AP}}}{\sqrt{a^2 + x_{ci}^2 + y_{cj}^2}}. \end{cases}$$

In practice, we create a mesh over a quarter prolate spheroid and obtain the full spheroid by reflecting the mesh along the planes $(X0Z)$ and $(X0Y)$ (see figure 1.7a).



(a) Division of the prolate spheroid into four subdomains. Here, the discretization is done with $n_x = 40$, $n_y = 20$, $n_{y1} = 20$.

(b) Full cubed spheroidal mesh obtained by reflecting the quarter spheroid along the planes $(X0Z)$ and $(X0Y)$.

Figure 1.7: Cubed spheroidal mesh.

1.2.2 Comparison of the two parametrizations

Each parametrization has its own advantages and inconveniences. We conducted a first analysis by computing the surface area of the prolate spheroid with the two discretizations and comparing the results to the known theoretical surface area.

Results show that the prolate spheroidal mesh is more precise by a full order of magnitude, irrespective of the total size of the mesh (i.e. total number of discretization points).

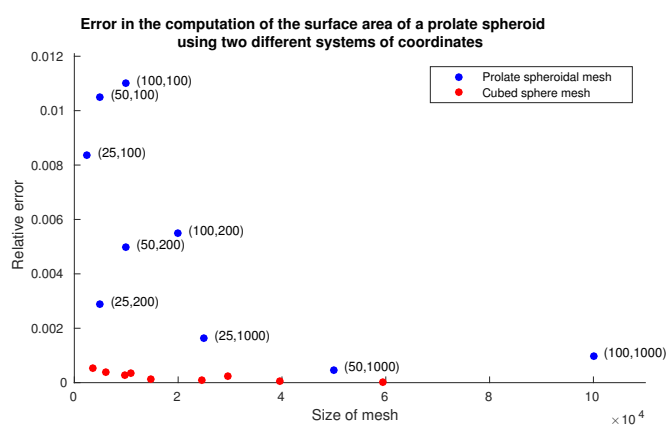


Figure 1.8: Comparison of the performance of the spheoridal mesh and the cubed spheroidal mesh in the computation of the total surface area of the prolate spheroid. Numbers in parentheses indicate the number of discretization points (N_θ, N_η).

1.2.3 Numerical approximation of the diffusion process

We compare the two parametrizations in the computation of the diffusion term. We chose to use finite differences method to approximate the diffusion process. We first explain the advantage of using the cubed spheroidal coordinates over the spheroidal ones.

Spheroidal coordinates

We recall that the CFL condition of the explicit scheme for diffusion on a rectangular domain of $\mathbb{R}^2$, with time-step Δt and spatial steps Δx and Δy is classically given by $D(\frac{D\Delta t}{\Delta x^2} + \frac{D\Delta t}{\Delta y^2}) \leq \frac{1}{2}$. Heuristically, this hints why the grid obtained with the spheroidal coordinates is ill-suited for the discretization of diffusion: it implies very small distance steps near the poles, thus requiring a very small time step Δt to be stable.

More rigorously, let us compare the CFL condition near the pole ($\eta = 0$) and near the equator ($\eta = \pi/2$).

Let $f^n_{i,j} := f(n\Delta t, \eta_i, \theta_j)$. From equation (1.7), we can compute the discrete Laplace-Beltrami operator using centered finite differences:

$$\begin{aligned}\Delta_{\mathrm{LB}} f^n_{i,j} \approx & \frac{1}{(h^0_\eta)_{i,j}(h^0_\theta)_{i,j}} \frac{1}{2\Delta\eta} \left(\frac{(h^0_\theta)_{i+1,j}}{(h^0_\eta)_{i+1,j}} \frac{f^n_{i+2,j} - f^n_{i,j}}{2\Delta\eta} - \frac{(h^0_\theta)_{i-1,j}}{(h^0_\eta)_{i-1,j}} \frac{f^n_{i,j} - f^n_{i-2,j}}{2\Delta\eta} \right) \\ & + \frac{1}{(h^0_\theta)^2_{i,j}} \frac{f^n_{i,j+2} - 2f^n_{i,j} + f^n_{i,j-2}}{4\Delta\theta^2}\end{aligned} \tag{1.26}$$

with

$$\begin{cases} (h^0_\eta)_{i,j} = a\sqrt{\sinh^2 \xi_0 + \sin^2 \eta_i} \\ (h^0_\theta)_{i,j} = a \sinh \xi_0 \sin \eta_i. \end{cases}$$

Notice that when $\eta_i \to 0$, $(h^0_\eta)_{i,j} \approx a \sinh \xi_0$ and $(h^0_\theta)_{i,j} \approx a \sinh \xi_0 \eta_i$. Hence, near the pole $\eta = 0$,

$$\begin{aligned}\Delta_{\mathrm{LB}} f^n_{i,j} \approx & \frac{1}{a^2 \sinh^2 \xi_0 \eta_i} \frac{1}{2\Delta\eta} \left(\eta_{i+1} \frac{f^n_{i+2,j} - f^n_{i,j}}{2\Delta\eta} - \eta_{i-1} \frac{f^n_{i,j} - f^n_{i-2,j}}{2\Delta\eta} \right) + \frac{f^n_{i,j+2} - 2f^n_{i,j} + f^n_{i,j-2}}{4(a \sinh \xi_0 \eta_i)^2 \Delta\theta^2} \\ \approx & \frac{1}{(a \sinh \xi_0 \; i\Delta\eta)^2} \frac{f^n_{i,j+2} - 2f^n_{i,j} + f^n_{i,j-2}}{4\Delta\theta^2}\end{aligned} \tag{1.27}$$

as the second term clearly dominates the first one.

Hence solving the heat equation numerically yields:

$$f_{i,j}^{n+1} \approx f_{i,j}^n + \Delta t \Delta_{\mathrm{LB}} f_{i,j}^n \approx f_{i,j}^n + \frac{\Delta t}{L_{\mathrm{DV}}^2\, i^2 4\Delta\eta^2\Delta\theta^2}(f_{i,j+2}^n - 2f_{i,j}^n + f_{i,j-2}^n) = (1-2\delta)f_{i,j}^n + \delta f_{i,j+2}^n + \delta f_{i,j-2}^n,$$

where $\delta := \frac{\Delta t}{L_{\mathrm{DV}}^2\, i^2 4\Delta\eta^2\Delta\theta^2}$. This means that $f_{i,j}^{n+1}$ is a convex combination of $f_{i,j+2}^n$, $f_{i,j}^n$ and $f_{i,j-2}^n$ if and only if $2\delta \leq 1$, i.e. if and only if:

$$\frac{\Delta t}{L_{\mathrm{DV}}^2\, i^2 4\Delta\eta^2\Delta\theta^2} \leq \frac{1}{2}. \tag{1.28}$$

This condition ensures that the maximum principle is satisfied, implying stability of the scheme.

Now, near the equator, i.e. when $\eta_i \to \pi/2$, notice that $(h_\eta^0)_{i,j} \approx a\sqrt{\sinh^2\xi_0 + 1}$ and $(h_\theta^0)_{i,j} \approx a\sinh\xi_0$. This implies:

$$\begin{aligned}\Delta_{\mathrm{LB}} f_{i,j}^n \approx& \frac{1}{(h_\eta^0)_{i,j}^2}\frac{1}{2\Delta\eta}\left(\frac{f_{i+2,j}^n - f_{i,j}^n}{2\Delta\eta} - \frac{f_{i,j}^n - f_{i-1,j}^n}{2\Delta\eta}\right) + \frac{1}{(h_\theta^0)_{i,j}^2}\frac{f_{i,j+1}^n - 2f_{i,j}^n + f_{i,j-1}^n}{4\Delta\theta^2}\\ \approx& \frac{1}{a^2(\sinh^2\xi_0 + 1)}\frac{f_{i+2,j}^n - 2f_{i,j}^n + f_{i-2,j}^n}{4\Delta\eta^2} + \frac{1}{a^2\sinh^2\xi_0}\frac{f_{i,j+1}^n - 2f_{i,j}^n + f_{i,j-1}^n}{4\Delta\theta^2}.\end{aligned}$$

Let $\delta_1 := \frac{\Delta t}{4a^2(\sinh^2\xi_0+1)\Delta\eta^2}$ and $\delta_2 := \frac{\Delta t}{4a^2\sinh^2\xi_0\Delta\theta^2}$. Solving the heat equation numerically yields:

$$\begin{aligned}f_{i,j}^{n+1} &\approx f_{i,j}^n + \Delta t\Delta_{\mathrm{LB}} f_{i,j}^n \approx f_{i,j}^n + \delta_1(f_{i+2,j}^n - 2f_{i,j}^n + f_{i-2,j}^n) + \delta_2(f_{i,j+2}^n - 2f_{i,j}^n + f_{i,j-2}^n)\\ &\approx (1 - 2\delta_1 - 2\delta_2)f_{i,j}^n + \delta_1 f_{i+2,j}^n + \delta_1 f_{i-2,j}^n + \delta_2 f_{i,j+2}^n + \delta_2 f_{i,j-2}^n.\end{aligned}$$

So $f_{i,j}^{n+1}$ is a convex combination of $f_{i,j}^n$, $f_{i+2,j}^n$, $f_{i-2,j}^n$, $f_{i,j+2}^n$ and $f_{i,j-2}^n$ if and only if $2\delta_1 + 2\delta_2 \leq 1$, i.e. if and only if:

$$\frac{\Delta t}{4a^2(\sinh^2\xi_0 + 1)\Delta\eta^2} + \frac{\Delta t}{4a^2\sinh^2\xi_0\Delta\theta^2} \leq \frac{1}{2}. \tag{1.29}$$

The CFL condition near the equator $\eta = \pi/2$ (1.29) is thus much less stringent than the one at the pole $\eta = 0$ (1.28). This shows that such an irregular mesh is not adapted to solving the heat equation on a spheroidal shape.

Cubed spheroidal coordinates

Using the cubed spheroidal parametrization, one can compute the metric tensor of S in each subdomain. Let $r = (x, y, z)$. Then in D_0, the metric tensor G_0 is given by:

$$G_0 = \begin{pmatrix} g_0^{11} & g_0^{12} \\ g_0^{21} & g_0^{22} \end{pmatrix},$$

with

$$\begin{cases} g_0^{11} := \|\partial_{x_c} r\|^2 = \frac{1}{(a^2+x_c^2+y_c^2)^3}(L_{\mathrm{DV}}^2(a^2+y_c^2)^2 + L_{\mathrm{DV}}^2 x_c^2 y_c^2 + L_{\mathrm{AP}}^2 a^2 x_c^2) \\ g_0^{22} := \|\partial_{y_c} r\|^2 = \frac{1}{(a^2+x_c^2+y_c^2)^3}(L_{\mathrm{DV}}^2 x_c^2 y_c^2 + L_{\mathrm{DV}}^2(a^2+x_c^2)^2 + L_{\mathrm{AP}}^2 a^2 y_c^2) \\ g_0^{12} = g_0^{21} := (\partial_{y_c} r) \cdot (\partial_{x_c} r) = \frac{1}{(a^2+x_c^2+y_c^2)^3}(-L_{\mathrm{DV}}^2(2a^2+x_c^2+y_c^2) + L_{\mathrm{AP}}^2 a^2 x_c^2 y_c^2). \end{cases}$$

Notice that the cubed spheroidal coordinates are not orthogonal, hence the metric tensor is not diagonal, unlike the metric tensor G_P calculated using the prolate spheroidal coordinates (1.5). Its expression is thus more complicated, but the coordinates do not have any singularity.

The expression of the metric tensor allows to compute the Laplace-Beltrami operator in the cubed spheroid coordinates in each domain. For example, in D_0,

$$\Delta_{\mathrm{LB}} f = \frac{1}{\sqrt{|g_0|}} \sum_{i=1}^{2} \partial_i \Big(\sum_{j=1}^{2} \sqrt{|g_0|} g_0^{ij} \partial_j f\Big). \tag{1.30}$$

Using cubed spheroidal coordinates requires subdividing the prolate spheroid into several subdomains, and treating the interfaces between domains with appropriate boundary conditions. Notice that as stated in (1.18), due to to the symmetry of the source function with respect to the $(X0Z)$ plane, the whole system is symmetric with respect to the $(X0Z)$ plane. It is then sufficient to solve numerically system (1.18) on a half prolate spheroid, and to recover the solution on the full domain by symmetry with respect to $(X0Z)$. The boundaries between domains (see Figure 1.7a) are treated with Dirichlet boundary conditions. The boundaries of the quarter spheroid inscribed in the $(X0Z)$ plane are treated with Neumann boundary conditions, for reasons of symmetry (see Figure 1.9).

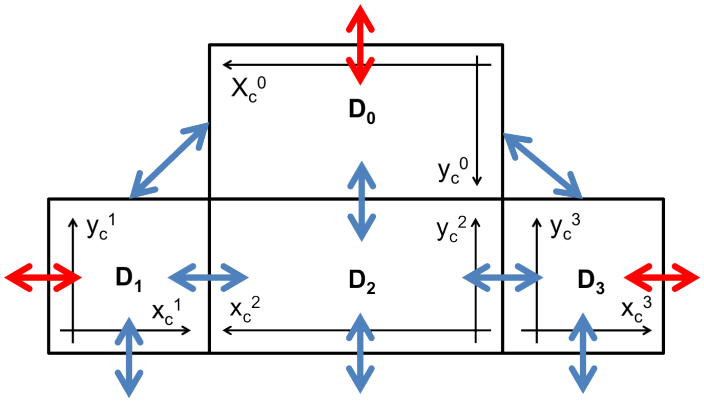
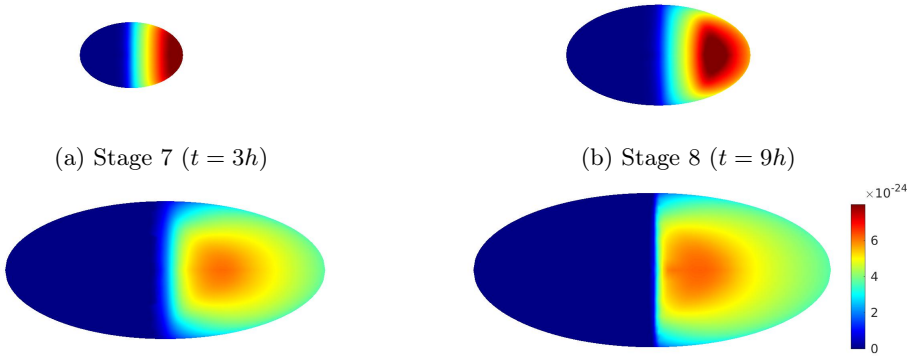


Figure 1.9: Interface conditions and boundary conditions for each subdomain of the quarter prolate spheroid. Blue: Dirichlet conditions. Red: Neumann boundary conditions.

1.3 Numerical results and experimental validation

Figure 1.10 shows the concentration of dpERK, the output of the model, at various stages of development. We tested our model by comparing its output (the spatio-temporal concentration of dpERK) with experimental measurements of the signal's intensity, with two different perturbations. Figure 1.11 shows the concentration of dpERK along the AP axis at stage 10A ($t = 21h$) in three different simulations. The wild-type simulation corresponds to the parameters defined in Table 1.1. The Sty RNAi perturbation corresponds to a modified γ_{Sty} for all $t > 6$, to cancel the effect of Sty on the dynamics after Stage 7. The EGFR RNAi perturbation corresponds to a reduced level of available receptors at the surface of the follicle cells. Several simulations were run to estimate the severity of the receptors reduction in the EGFR RNAi perturbation. Indeed, the exact proportion by which they are depleted is unknown. However, we are able to measure the resulting intensity of the signal, and working backwards we are able to estimate the amount of available receptors.



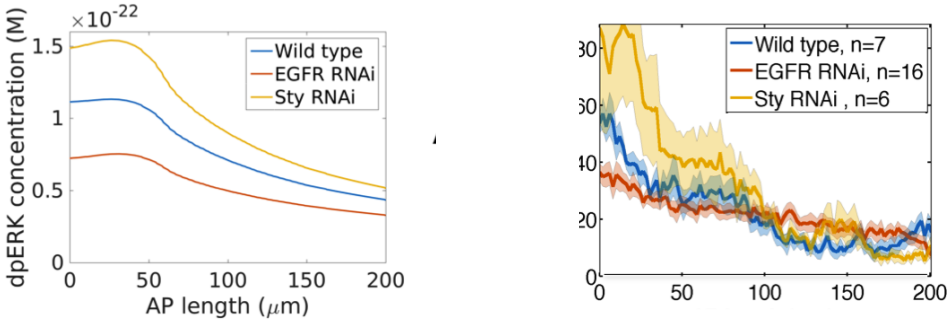
(a) Stage 7 ($t = 3h$)

(b) Stage 8 ($t = 9h$)

(c) Stage 9 ($t = 15h$)

(d) Stage 10A ($t = 21h$)

Figure 1.10: Numerical results: concentration of dpERK (in $mol/\mu m^2$) at four different times.



(a) dpERK concentration at Stage 10A along the anterior-posterior axis

(b) Experimental measurements of dpERK concentration (arbitrary units)

Figure 1.11: Comparison of simulation results to experimental measurements of dpERK intensity.

Chapter 2

Developmental Partial Differential Equations

Our work in developing a model for the spatio-temporal evolution of the Gurken morphogen (the initiator of the EGFR signaling pathway) in *Drosophila melanogaster* (see Chapter 1) has led us to identify the need for a suitable mathematical framework for reaction-diffusion equations in developing organisms.

Modeling the growth of living organism attracted the interest of many investigators both in the field of Developmental Biology and in the field of Applied Mathematics. Developmental biologists have shown that development is primarily induced by morphogens, which act on the organism as signals by triggering signaling pathways and provoking a response resulting in cell growth or differentiation [133]. Several modeling approaches have been explored from the mathematical point of view. From a microscopic standpoint, tissues are considered as a collection of cells, and discrete models such as cellular automata are used. We instead adopt a macroscopic standpoint, where the relevant quantity is the density of the signal on a manifold.

As seen in Chapter 1, Gurken diffuses in a thin space, called perivitelline space, which can be modeled by an evolving surface. This leads naturally to model the growing organism by coupling a growing surface with a signal diffusing on it, see [100]. Because of the biological motivation, this framework was called *Developmental Partial Differential Equations*.

We consider a general model, where the boundary of the organism is described by a Riemannian manifold, that evolves with respect to time due to the growth induced by the signal on it. In turn the evolution (for instance, heat diffusion) of the signal on the manifold is affected by the shape of the manifold. Indeed, intrinsic heat diffusion is described by the heat equation with the Laplace-Beltrami operator. Our aim is to investigate the coupling between growth and diffusion. There is a wide literature of studies for PDEs on manifolds, see for instance [113, 120], or Turing Patterns on evolving manifolds, see for instance [5, 76]. However the coupling of PDE and time-evolving

manifold was newly introduced in [100].

As a first step to understanding what shapes of the manifold can be attained from an initial configuration, we explore the non-commutativity of the growth (manifold change in time) and the diffusion operator (on the manifold itself). A newly defined concept of Lie bracket between the diffusion (2nd order operator) and growth (1st order operator) is able to capture such non-commutativity and thus provide new shapes towards which the manifold may evolve. As in classical geometric control theory [1, 13, 117], the concept of Lie bracket may indeed enclose all the needed information to capture the controlled dynamics. Moreover, such bracket can be understood as a new available direction for the growth of the organism.

We begin by introducing the general model, or Developmental Partial Differential Equation (DPDE) describing the coupling of growth and diffusion on a Riemannian manifold. We then prove existence and uniqueness of the solution to the DPDE by introducing a numerical scheme that discretizes time and solves diffusion and growth independently on each time interval. We prove that the limit of the scheme is the solution to the DPDE. We then use the scheme to define a new kind of Lie bracket between the diffusion and the growth operators. By computing the bracket explicitly, we show that it is not zero. Numerical simulations confirm the analytical computation of the bracket. Lastly, we study the control of a simplified problem: leveraging on natural symmetries of the egg chamber we choose as $\mathcal{M}$ a one-dimensional symmetric manifold embedded in $\mathbb{R}^2$ and initially equal to $\mathbb{S}^1$. Our main aim is to show controllability in terms of the possible shapes reachable from $\mathbb{S}^1$ regulating one or more sources. We show how to adapt the approach of Laroche, Martin and Rouchon [69], proving flatness of the heat equation, to our setting and then provide numerical studies.

2.1 The heat equation on time-varying manifolds

Let $\mathcal{P}_c(\mathbb{R}^d)$ denote the space of probability measures in $\mathbb{R}^d$ with compact support. We endow this space of measures with the weak topology of measures. Then, for example, we write $\lim_{n\to\infty} \mu^n = \mu^*$ to denote $\mu^n \rightharpoonup \mu^*$ when $n \to \infty$. Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$. Lastly, let $\mathcal{P}(\mathcal{M})$ denote the space of probability measures on $\mathcal{M}$, endowed with the Wasserstein distance W_p, whose definition is recalled below (see Definition 2.1.1).

2.1.1 The heat and transport evolutions

Let $\mathcal{M}_t$ be a time-evolving compact manifold embedded in $\mathbb{R}^d$, endowed with the Riemannian structure induced by the embedding. Let $\mu_t \in \mathcal{P}(\mathcal{M}_t)$ be a probability measure on $\mathcal{M}_t$. Remark

that since $\mathcal{M}_t$ is embedded in $\mathbb{R}^d$, μ_t can also be considered as a probability measure with compact support in $\mathbb{R}^d$, i.e. $\mu_t \in \mathcal{P}_c(\mathbb{R}^d)$. Both points of view will be useful. The manifold $\mathcal{M}_t$ evolves according to a vector field depending on the measure itself, $v[\cdot] : \mathcal{P}_c(\mathbb{R}^d) \to \mathrm{Lip}(\mathbb{R}^d, \mathbb{R}^d) \cap \mathcal{L}^\infty(\mathbb{R}^d)$, with the following assumptions on v:

- $v[\mu]$ is uniformly bounded, i.e. there exists $M > 0$ such that for all $\mu \in \mathcal{P}_c(\mathbb{R}^d)$, for all $x \in \mathbb{R}^d$, $|v[\mu](x)| \leq M$
- $v[\mu]$ is uniformly Lipschitz, i.e. there exists $L > 0$ such that for all $\mu \in \mathcal{P}_c(\mathbb{R}^d)$, for all $x, y \in \mathbb{R}^d$, $|v[\mu](x) - v[\mu](y)| \leq L|x - y|$
- v is a Lipshitz function, i.e. there exists K such that for all $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$, $\|v[\mu] - v[\nu]\|_{C^0} \leq KW_2(\mu, \nu)$

In the third assumption, $W_2(\mu, \nu)$ denotes the 2-Wasserstein distance between μ and ν. We recall the definition of the Wasserstein distance (see [129]). For every probability measure μ and measurable map ϕ, the push-forward $\phi\#\mu$ is defined by $\phi\#\mu(A) = \mu(\phi^{-1}(A))$.

Definition 2.1.1. *Let $p \geq 1$. Given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the p-Wasserstein distance between μ and ν is given by:*

$$\mathcal{W}_p(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p d\pi(x, y) \right)^{1/p}$$

where $\Pi(\mu, \nu)$ is the set of transference plans from μ to ν, i.e. of the probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ, ν, respectively. In other words $P_x\#\pi = \mu$ and $P_y\#\pi = \nu$ (where P_x, respectively P_y denote the projection on the first, respectively second, component of (x, y).)

The transference plans in $\Pi(\mu, \nu)$ can be seen as methods to transport μ to ν and the term $\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p d\pi(x, y)$ can be interpreted as a cost (as p-power of the distance) to move the mass of μ onto the mass of ν via the plan π. Hence, the Wasserstein distance is the minimal cost to move one mass over the other. For a complete introduction to the topic of Wasserstein distances we refer the reader to [129].

For each time $t \in [0, T]$, the manifold $\mathcal{M}_t$ is endowed with the following elements:

- the volume form dm_t, given by the Riemannian structure;
- the intrinsic Laplacian Δ_t (also called the Laplace-Beltrami operator), that is intrinsically defined by the Riemannian structure;

- the heat kernel $p^{\tau}_{\mathcal{M}_t}(x,y)$ that provides the solution of the heat equation $\partial_\tau \mu = \Delta_t \mu$ with initial data $\mu(0,x)$ by convolution as follows:

$$\mu(\tau, x) = \int_{\mathcal{M}_t} p^{\tau}_{\mathcal{M}_t}(x,y)\, d\mu(0,y).$$

We also denote the solution of the heat equation on $\mathcal{M}_t$ by the following semigroup notation:

$$\mu(\tau) = e^{\tau \Delta_t} \mu(0).$$

Remark that all these elements are defined for a fixed t, hence for a fixed manifold $\mathcal{M}_t$.

The measure μ_t is affected by the evolution of the manifold via the vector field $v[\mu_t]$ and by the diffusion via the Laplace-Beltrami operator Δ_t. The combination of these two phenomena give the evolution of μ_t through the following *Developmental Partial Differential Equation*:

$$\partial_t \mu_t + \nabla \cdot (v[\mu_t]\mu_t) = \Delta_t \mu_t, \tag{2.1}$$

where the manifold $\mathcal{M}_t$ is the support of μ_t at each time $t > 0$. Since μ_t are measures in $\mathbb{R}^d$, such equation needs to be interpreted in the weak sense, i.e. for all $f \in C^\infty(\mathbb{R}^d)$ it holds

$$\partial_t \int_{\mathbb{R}^d} f d\mu_t - \int_{\mathbb{R}^d} (\nabla f \cdot v[\mu_t]) d\mu_t = \int_{\mathcal{M}_t} \Delta_t f \; d\mu_t. \tag{2.2}$$

We prove existence and uniqueness of a solution to Equation (2.1).

Theorem 2.1.1. *Let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ be a probability measure with compact support in $\mathbb{R}^d$. Let $\mathcal{M}_0$ denote the support of μ_0, and suppose that $\mathcal{M}_0$ is a compact Riemannian manifold embedded in $\mathbb{R}^d$. Let $v[\cdot] : \mathcal{P}_c(\mathbb{R}^d) \to \mathrm{Lip}(\mathbb{R}^d, \mathbb{R}^d) \cap \mathcal{L}^\infty(\mathbb{R}^d)$, satisfy the following assumptions:*

- *there exists $M > 0$ such that for all $\mu \in \mathcal{P}_c(\mathbb{R}^d)$, for all $x \in \mathbb{R}^d$, $|v[\mu](x)| \leq M$*
- *there exists $L > 0$ such that for all $\mu \in \mathcal{P}_c(\mathbb{R}^d)$, for all $x, y \in \mathbb{R}^d$, $|v[\mu](x) - v[\mu](y)| \leq L|x-y|$*
- *there exists K such that for all $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$, $\|v[\mu] - v[\nu]\|_{C^0} \leq K W_2(\mu, \nu)$.*

Then there exists a unique solution to the developmental partial differential equation (2.1)*:*

$$\partial_t \mu_t + \nabla \cdot (v[\mu_t]\mu_t) = \Delta_t \mu_t,$$

with initial data μ_0, where Δ_t denotes the Laplace-Beltrami operator of $\mathcal{M}_t$, the support of μ_t.

We prove existence of a solution to (2.1) as the limit of a numerical scheme performing alternating steps of diffusion and transport for fixed time intervals. The idea of defining the solution as the limit of a discrete scheme was introduced in [95, 96] for the transport equation:

$$\partial \mu_t + \nabla \cdot (v[\mu_t]\mu_t) = 0. \tag{2.3}$$

The main difference between the transport equation 2.3 and our developmental PDE (2.1) is the diffusion term $\Delta_t \mu_t$. The idea of the proof lies in decoupling the diffusion and transport phenomena.

More specifically, let $T \in \mathbb{R}$ be the final time and let $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ denote an initial compactly supported measure. For a given discretization parameter $n \in \mathbb{N}$, we define a sequence of curves (μ_s^n) via the following scheme:

Scheme $\mathbb{S}$

Define $\tau_n = t_n := 2^{-n}T$. Let $\mu^n(0) := \mu_0$, and $\mathcal{M}_0^n := \mathcal{M}_0$. On the nodes lt_n (with $l \in \{0, ..., 2^n - 1\}$) we define $\mu^n((l+1)t_n)$ from $\mu^n(lt_n)$ as follows:

1. Define $\tilde{\mu}^n(lt_n) := e^{\Delta^n_{lt_n}\tau_n}(\mu^n(lt_n))$, i.e the solution of the heat equation on $\mathcal{M}^n_{lt_n}$ with initial data $\mu^n(lt_n)$ at time τ_n.

2. Let $\phi_t^{n,l}$ be the flow of $v[\tilde{\mu}^n(lt_n)]$ and let $\mu^n((l+1)t_n) := \phi^{n,l}_{t_n}\#\tilde{\mu}^n(lt_n)$, i.e. the push-forward of $\tilde{\mu}^n(lt_n)$ via the flow $\phi^{n,l}_{t_n}$. We define: $\mathcal{M}^n_{(l+1)t_n} := \phi^{n,l}_{t_n}\#\mathcal{M}^n_{lt_n}$.

In between nodes, for $s \in [0, t_n/2]$, we define $\mu^n(lt_n + t) := e^{\Delta^n_{lt_n}2s}(\mu^n(lt_n))$, a measure on $\mathcal{M}^n_{lt_n}$. We define $\mu^n((l+\frac{1}{2})t_n + s) = \phi^{n,l}_{2s}\#(e^{\Delta^n_{lt_n}t_n}(\mu^n(lt_n)))$, a measure on $\mathcal{M}^n_{(l+\frac{1}{2})t_n+s} := \phi^{n,l}_{2s}\#\mathcal{M}^n_{lt_n}$.

In the definition of $\mathbb{S}$, we distinguish t_n and τ_n for better description and approximation of the two phenomena of deformation and heat diffusion. This is only for clarity and in reality we will study the limit of the scheme when $n \to \infty$ for $t_n = \tau_n$.

We first give properties regarding the commutativity of the push-forward and heat operators. Indeed in order to prove convergence of the scheme $\mathbb{S}$ we need an estimate of the error made when performing a step of diffusion followed by transport compared to performing a step of transport followed by diffusion. Lemma 2.1.2 provide this estimate for a given vector field v independant of μ.

Lemma 2.1.1. *Let $\mu_0 \in \mathcal{P}(\mathcal{M})$ and $t > 0$. Let $v : R^d \to \mathbb{R}^d$ be a vector field. Let δ_x denote the Dirac measure at $x \in \mathcal{M}$. Let $p^t(\cdot,\cdot) : \mathcal{M} \times \mathcal{M} \mapsto \mathcal{P}(\mathcal{M} \times \mathcal{M})$ denote the heat kernel on $\mathcal{M}$ and*

$\tilde{p}^t(\cdot,\cdot) : \tilde{M} \times \tilde{M} \mapsto \mathcal{P}(\tilde{M} \times \tilde{M})$ *the heat kernel on the pushforward* $\tilde{\mathcal{M}}$ *of* $\mathcal{M}$ *via* ϕ_v^t. *There exists* $x \in \mathcal{M}$ *such that*

$$\mathcal{W}_2(\phi_v^t\#(p^t * \mu_0), \tilde{p}^t * (\phi_v^t\#\mu_0)) \leq \mathcal{W}_2(\phi_v^t\#(p^t * \delta_x), \tilde{p}^t * (\phi_v^t\#\delta_x)).$$

Proof. Let $\mu_0 \in \mathcal{P}(\mathcal{M})$. Let $F : \mu \mapsto \mathcal{W}_2(\phi_v^t\#(p^t * \mu), \tilde{p}^t * (\phi_v^t\#\mu))$. F is a continuous function from $\mathcal{P}(\mathcal{M})$ to $\mathbb{R}$. We construct a sequence $(\bar{\mu}^k)_{k\in\mathbb{N}}$ such that $\bar{\mu}^0 = \mu^0$, and for all $k \in \mathbb{N}$, $F(\bar{\mu}^k) \leq F(\bar{\mu}^{k+1})$ and $\mathrm{diam}(\mathrm{supp}(\bar{\mu}^k)) \xrightarrow[k\to\infty]{} 0$. Let $k \in \mathbb{N}$ and $\epsilon_k = \frac{1}{2^k}\mathrm{diam}(\mathrm{supp}(\mu_0))$. By compactness of $\mathrm{supp}(\bar{\mu}^k)$, there exists a finite set $\{x_i^k\}_{i\in\{1,\dots,N_k\}}$ such that $\mathrm{supp}(\bar{\mu}^k) \subset \cup_{i=1}^{N_k} B(x_i^k, \epsilon_k)$, where we denote by $B(x,r)$ the geodesic ball centered at $x \in \mathcal{M}$ and of radius r. Define $\nu_1^k = \bar{\mu}^k_{|B(x_1^k,\epsilon_k)}$ and for all $i \in \{1,\dots,N_k\}$, $\nu_i^k = \bar{\mu}^k_{|B(x_i^k,\epsilon_k)} - \sum_{j=1}^{i-1} \nu_{j\,|B(x_i^k,\epsilon_k)}^k$. Let $\bar{\nu}_i^k = \frac{\nu_i^k}{|\nu_i^k|}$ and $\lambda_i^k = |\nu_i^k|$. Then $\bar{\mu}^k = \sum_{i=1}^{N_k} \lambda_i^k \bar{\nu}_i^k$, and for all $i \in \{1,\dots,N_k\}$, $\lambda_i^k \in [0,1]$ and $\sum_{i=1}^{N_k} \lambda_i^k = 1$.

$$\begin{aligned} F(\bar{\mu}^k) &= \mathcal{W}_2(\phi_v^t\#(p^t * \bar{\mu}^k), \tilde{p}^t * (\phi_v^t\#\bar{\mu}^k)) = \mathcal{W}_2(\phi_v^t\#(p^t * \sum_{i=1}^{N_k} \lambda_i^k \bar{\nu}_i^k), \tilde{p}^t * (\phi_v^t\# \sum_{i=1}^{N_k} \lambda_i^k \bar{\nu}_i^k))) \\ &\leq \sum_{i=1}^{N_k} \lambda_i^k \mathcal{W}_2(\phi_v^t\#(p^t * \bar{\nu}_i^k), \tilde{p}^t * (\phi_v^t\#\bar{\nu}_i^k))) \leq \max_{i\in\{1,\dots,N_k\}} \mathcal{W}_2(\phi_v^t\#(p^t * \bar{\nu}_i^k), \tilde{p}^t * (\phi_v^t\#\bar{\nu}_i^k))). \end{aligned}$$

Let $m := \arg\max_{i\in\{1,\dots,N_k\}} \mathcal{W}_2(\phi_v^t\#(p^t * \bar{\nu}_i^k), \tilde{p}^t * (\phi_v^t\#\bar{\nu}_i^k)))$ and define $\bar{\mu}^{k+1} := \bar{\nu}_m^k$. Then $F(\bar{\mu}^k)) \leq F(\bar{\mu}^{k+1})$ and $\mathrm{diam}(\mathrm{supp}(\bar{\mu}^{k+1})) = \epsilon_k = \frac{1}{2^k}\mathrm{diam}(\mathrm{supp}(\mu_0))$. The support of the constructed sequence $(\bar{\mu}^k)$ tends to a single point, while for all $k \in \mathbb{N}$, $|\bar{\mu}^k| = 1$. Hence there exists $x \in \mathcal{M}$ such that $\lim_{k\to\infty} \bar{\mu}^k = \delta_x$. By continuity of F, $F(\mu_0) \leq F(\bar{\mu}^1) \leq F(\bar{\mu}^2) \dots \leq F(\delta_x)$. □

We now use Lemma 2.1.1 to bound from above the Wasserstein distance between the transport of the convolution of a measure with the heat kernel and the convolution of its transport with the heat kernel.

Lemma 2.1.2. *Let* $\mu_0 \in \mathcal{P}(\mathcal{M})$ *and* $t > 0$. *Let* $v \in \mathrm{Lip}(\mathbb{R}^d, \mathbb{R}^d)$ *be a vector field. As previously, let* $p^t(\cdot,\cdot) : \mathcal{M} \times \mathcal{M} \mapsto \mathcal{P}(\mathcal{M} \times \mathcal{M})$ *denote the heat kernel on* $\mathcal{M}$ *and* $\tilde{p}^t(\cdot,\cdot) : \tilde{M} \times \tilde{M} \mapsto \mathcal{P}(\tilde{M} \times \tilde{M})$ *the heat kernel on the pushforward* $\tilde{\mathcal{M}}$ *of* $\mathcal{M}$ *via* ϕ_v^t. *Then there exists a constant* β *such that for* t *small enough,*

$$\mathcal{W}_2(\phi_v^t\#(p^t * \mu_0), \tilde{p}^t * (\phi_v^t\#\mu_0)) \leq \beta t\sqrt{t}.$$

Remark 2.1.1. As a first approach, we give a proof of Lemma 2.1.2 in the simplified case: $\mathcal{M} = \mathbb{R}^n$. The generalization to any Riemannian manifold will require additional assumptions on the curvature of $\mathcal{M}$.

Proof. From Lemma 2.1.1, it is sufficient to prove that for all $x \in \mathcal{M}$,

$$\mathcal{W}_2(\phi_v^t\#(p^t * \delta_x), \tilde{p}^t * (\phi_v^t\#\delta_x)) \leq \beta t\sqrt{t}.$$

Let $x \in M$. We first suppose that $\mathcal{M} = \tilde{\mathcal{M}} = \mathbb{R}^n$. Let $y \in \mathcal{M}$. The vector field evaluated at y satisfies:

$$v(y) = v(x) + Jv(x) \cdot (y - x) + O(\|y - x\|^2)$$

where $Jv(x)$ denotes the Jacobian of v at x. Let $\tilde{x} = \phi_v^t(x)$ and $\tilde{y} = \phi_v^t(y)$. We have:

$$\tilde{y} = \phi_v^t(y) = y + tv(y) + O(t^2) = y + t[v(x) + Jv(x) \cdot (y - x) + O(\|y - x\|^2)] + O(t^2).$$

Hence, denoting by I the identity matrix, $\tilde{y} - \tilde{x} = (I + tJv(x)) \cdot (y - x) + t\, O(\|y - x\|^2) + O(t^2)$, so $y - x = (I + tJv(x))^{-1}(\tilde{y} - \tilde{x} + t\, O(\|\tilde{y} - \tilde{x}\|^2) + O(t^2))$, which can be rewritten as:

$$y - x = (I - tJv(x))(\tilde{y} - \tilde{x}) + O(t)O(\|\tilde{y} - \tilde{x}\|^2) + O(t^2).$$

Let $\mu_1 := \phi_v^t\#(p^t * \delta_x)$ and $\mu_2 := \tilde{p}^t * (\phi_v^t\#\delta_x)$. By definition of the push-forward and of the heat kernel,

$$\mu_2 = \tilde{p}^t * \delta_{\phi_v^t(x)} = \tilde{p}^t(\tilde{x}, \cdot)$$

i.e.

$$\mu_2(\tilde{y}) = \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{\|\tilde{y} - \tilde{x}\|^2}{4t}\right).$$

On the other hand, $\forall A \subset \tilde{M}$,

$$\begin{aligned}\mu_1(A) &= \int_A \phi_v^t\#(p^t * \delta_x) = (p^t * \delta_x)(\phi_v^{-t}(A)) = \int_{\phi_v^{-t}(A)} p^t(x, y)dy \\ &= \int_{\phi_v^{-t}(A)} \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{\|y - x\|^2}{4t}\right) dy.\end{aligned}$$

By change of variable $y \mapsto \tilde{y}$, with: $y = \phi_v^{-t}(\tilde{y})$, $dy = |\det(I - tJv(x))| d\tilde{y}$,

$$\begin{aligned}
\mu_1(A) &= \int_A \frac{|\det(I - tJv(x))|}{(4\pi t)^{n/2}} \exp\left(-\frac{\|(I - tJv(x))(\tilde{y} - \tilde{x}) + O(t)O(\|\tilde{y} - \tilde{x}\|^2) + O(t^2)\|^2}{4t}\right) d\tilde{y} \\
&= \int_A \frac{|\det(I - tJv(x))|}{(4\pi t)^{n/2}} \exp\left(-\frac{\|(I - tJv(x))(\tilde{y} - \tilde{x})\|^2}{4t}\right) \\
&\quad \exp\left(-\frac{(O(t)O(\|\tilde{y} - \tilde{x}\|^2) + O(t^2))^2 + (1 + O(t))O(\|\tilde{y} - \tilde{x}\|)(O(t)O(\|\tilde{y} - \tilde{x}\|^2) + O(t^2))}{4t}\right) d\tilde{y} \\
&= \int_A \frac{|\det(I - tJv(x))|}{(4\pi t)^{n/2}} \exp\left(-\frac{\|(I - tJv(x))(\tilde{y} - \tilde{x})\|^2}{4t}\right) \\
&\qquad \exp\left(O(t^3) + O(t)O(\|\tilde{y} - \tilde{x}\|^2) + O(t^2)O(\|\tilde{y} - \tilde{x}\|)\right) d\tilde{y} \\
&= \int_A \frac{|\det(I - tJv(x))|}{(4\pi t)^{n/2}} \exp\left(-\frac{\|(I - tJv(x))(\tilde{y} - \tilde{x})\|^2}{4t}\right) \\
&\qquad \left(1 + O(t^3) + O(t^2)O(\|\tilde{y} - \tilde{x}\|) + O(t)O(\|\tilde{y} - \tilde{x}\|^2)\right) d\tilde{y}.
\end{aligned}$$

Hence μ_2 is a Gaussian of covariance matrix $\Sigma_2 = (\frac{1}{2t}I)^{-1} = 2tI$ centered at $\tilde{x}$, while μ_1 is a perturbed Gaussian of covariance matrix $\Sigma_1 = (\frac{1}{2t}(I - tJv(x))^T(I - tJv(x)))^{-1} = 2t((I - tJv(x))^T(I - tJv(x)))^{-1}$ centered at $\tilde{x}$. Let us denote by $\mathcal{N}_1$ the Gaussian centered at $\tilde{x}$ of variance Σ_1, and by $\mathcal{N}_2$ the Gaussian centered at $\tilde{x}$ of variance Σ_2. Then μ_1 and μ_2 write:

$$\mu_1 = \mathcal{N}_1 + \rho_1 \quad \text{and} \quad \mu_2 = \mathcal{N}_2,$$

with:

$$\begin{cases}
\mathcal{N}_1(\tilde{y}) = \frac{|\det(I - tJv(x))|}{(4\pi t)^{n/2}} \exp\left(-\frac{\|(I - tJv(x))(\tilde{y} - \tilde{x})\|^2}{4t}\right) \\
\rho_1(\tilde{y}) = (O(t^3) + O(t^2)O(\|\tilde{y} - \tilde{x}\|) + O(t)O(\|\tilde{y} - \tilde{x}\|^2))\mathcal{N}_1(\tilde{y}) \\
\mathcal{N}_2(\tilde{y}) = \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{\|\tilde{y} - \tilde{x}\|^2}{4t}\right).
\end{cases}$$

We now estimate the Wasserstein distance between μ_1 and μ_2. Let $R > 0$. We denote by μ^R the restriction of μ to the ball centered at $\tilde{x}$ of radius R, i.e. $\mu^R := \mu|_{B(\tilde{x},R)}$. Notice that μ_1 and μ_2 are probability measures, hence of the same mass. This is not the case with μ_1^R and μ_2^R. Since we will deal with measures of different masses, we use the generalized Wasserstein distance $\mathcal{W}_2^{a,b}$ (see [96] for a definition). From the triangular inequality we write:

$$\mathcal{W}_2^{a,b}(\mu_1, \mu_2) \leq \mathcal{W}_2^{a,b}(\mu_1, \mu_1^R) + \mathcal{W}_2^{a,b}(\mu_1^R, \mathcal{N}_1^R) + \mathcal{W}_2^{a,b}(\mathcal{N}_1^R, \mathcal{N}_1) + \mathcal{W}_2^{a,b}(\mathcal{N}_1, \mathcal{N}_2). \tag{2.4}$$

Let us study the first term of (2.4). Notice that

$$\mathcal{W}_2^{a,b}(\mu_1,\mu_1^R)=\mathcal{W}_2^{a,b}(\mu_1^R+(\mu_1-\mu_1^R),\mu_1^R+0)\le\mathcal{W}_2^{a,b}(\mu_1^R,\mu_1^R)+\mathcal{W}_2^{a,b}(\mu_1-\mu_1^R,0)\le a|\mu_1-\mu_1^R|$$

where the last inequality is a direct application of Proposition 2 in [96]. Hence we are left with the task of estimating the mass of the "tail" of the perturbed Gaussian μ_1:

$$|\mu_1-\mu_1^R|:=1-\int_{\|\tilde{y}-\tilde{x}\|\le R}\mu_1(\tilde{y})d\tilde{y}.$$

One can prove that for a Gaussian defined by a positive-definite covariance matrix A,

$$\int_{\|x\|\le R}e^{-x^TA^TAx}dx\ge\frac{1}{\det(A)}(\pi(1-e^{-nR\lambda_m(A)}))^{n/2}$$

where $\lambda_m(A)$ denotes the smallest eigenvalue of A. Applied to $\mathcal{N}_1$, we have:

$$\begin{aligned}\int_{\|\tilde{y}-\tilde{x}\|\le R}\frac{|\det(I-tJv(x))|}{(4\pi t)^{n/2}}\exp\left(-\frac{\|(I-tJv(x))(\tilde{y}-\tilde{x})\|^2}{4t}\right)d\tilde{y}\\ \ge\frac{|\det(I-tJv(x))|}{(4\pi t)^{n/2}}\frac{(\pi(1-e^{-nR\lambda_m}))^{n/2}}{\det(I-tJv(x))}.\end{aligned}$$

where λ_m denotes the smallest eigenvalue of $I-tJv(x)$. For t small enough, $\det(I-tJv(x))>0$, so

$$\int_{\|\tilde{y}-\tilde{x}\|\le R}\frac{|\det(I-tJv(x))|}{(4\pi t)^{n/2}}\exp\left(-\frac{\|(I-tJv(x))(\tilde{y}-\tilde{x})\|^2}{4t}\right)d\tilde{y}\ge\left(\frac{1-e^{-nR\lambda_m}}{4t}\right)^{n/2}.$$

Let $R=-\frac{1}{n\lambda_m}\ln\left(1-4t(1-t\sqrt{t})^{-n/2}\right)$. Then

$$\int_{\|\tilde{y}-\tilde{x}\|\le R}\frac{|\det(I-tJv(x))|}{(4\pi t)^{n/2}}\exp\left(-\frac{\|(I-tJv(x))(\tilde{y}-\tilde{x})\|^2}{4t}\right)d\tilde{y}\ge1-t\sqrt{t}. \tag{2.5}$$

We now compute

$$\begin{aligned}\int_{\|\tilde{y}-\tilde{x}\|\le R}\mu_1(\tilde{y})d\tilde{y}&=\int_{\|\tilde{y}-\tilde{x}\|\le R}\mathcal{N}_1(\tilde{y})(1+O(t^3)+O(t^2)O(\|\tilde{y}-\tilde{x}\|)+O(t)O(\|\tilde{y}-\tilde{x}\|^2))d\tilde{y}\\&=\int_{\|\tilde{y}-\tilde{x}\|\le R}\mathcal{N}_1(\tilde{y})d\tilde{y}(1+O(t^3)+O(t^2)O(R)+O(t)O(R^2)).\end{aligned}$$

Notice that for t small enough, $R=-\frac{1}{n\lambda_m}\ln\left(1-4t(1-\frac{n}{2}t\sqrt{t}+O(t^3))\right)=O(t)$. Then

$$\int_{\|\tilde{y}-\tilde{x}\|\le R}\mu_1(\tilde{y})d\tilde{y}=\int_{\|\tilde{y}-\tilde{x}\|\le R}\mathcal{N}_1(\tilde{y})d\tilde{y}(1+O(t^3)+O(t^2)O(t)+O(t)O(t^2))\ge1-t\sqrt{t}+O(t^3).$$

This means that we can estimate the first term of (2.4) as follows:

$$\mathcal{W}_2^{a,b}(\mu_1, \mu_1^R) \leq at\sqrt{t}. \tag{2.6}$$

We now look at the second term of (2.4).

$$\mathcal{W}_2^{a,b}(\mu_1^R, \mathcal{N}_1^R) = \mathcal{W}_2^{a,b}(\mathcal{N}_1^R + \rho_1^R, \mathcal{N}_1^R) \leq \mathcal{W}_2^{a,b}(\mathcal{N}_1^R, \mathcal{N}_1^R) + \mathcal{W}_2^{a,b}(\rho_1^R, 0) \leq a|\rho_1^R|$$

where again the last inequality comes from Proposition 2 of [96]. We then compute:

$$\begin{aligned} |\rho_1^R| &= \int_{\|\tilde{y}-\tilde{x}\|\leq R} \mathcal{N}_1(\tilde{y})(O(t^3) + O(t^2)O(\|\tilde{y}-\tilde{x}\|) + O(t)O(\|\tilde{y}-\tilde{x}\|^2))d\tilde{y} \\ &= \int_{\|\tilde{y}-\tilde{x}\|\leq R} \mathcal{N}_1(\tilde{y})d\tilde{y}O(t^3) = O(t^3). \end{aligned}$$

Hence we estimated:

$$\mathcal{W}_2^{a,b}(\mu_1^R, \mathcal{N}_1^R) \leq O(t^3). \tag{2.7}$$

Now the third term of (2.4) can be estimated by the mass of the tail of the Gaussian $\mathcal{N}_1$, as follows:

$$\mathcal{W}_2^{a,b}(\mathcal{N}_1^R, \mathcal{N}_1) \leq \mathcal{W}_2^{a,b}(\mathcal{N}_1^R, \mathcal{N}_1^R) + \mathcal{W}_2^{a,b}(0, \mathcal{N}_1 - \mathcal{N}_1^R) \leq a|\mathcal{N}_1 - \mathcal{N}_1^R|$$

From the estimate (2.5), $|\mathcal{N}_1 - \mathcal{N}_1^R| = 1 - |\mathcal{N}_1^R| \leq t\sqrt{t}$. So the third term of (2.4) becomes:

$$\mathcal{W}_2^{a,b}(\mathcal{N}_1^R, \mathcal{N}_1) \leq at\sqrt{t}. \tag{2.8}$$

Lastly, we are left with estimating the Wasserstein distance between two Gaussians $\mathcal{N}_1$ and $\mathcal{N}_2$ centered at $\tilde{x}$ and of covariance matrices $\Sigma_1 = 2t((I - tJv(x))^T(I - tJv(x)))^{-1}$ and $\Sigma_2 = 2tI$. From [44], we have:

$$\mathcal{W}_2(\mathcal{N}_1, \mathcal{N}_2)^2 = \|\tilde{x} - \tilde{x}\|_2^2 + \mathrm{Tr}\,(\Sigma_1) + \mathrm{Tr}\,(\Sigma_2) - 2\mathrm{Tr}\left([\sqrt{\Sigma_1}\Sigma_2\sqrt{\Sigma_1}]^{1/2}\right).$$

For t small enough,

$$\begin{aligned} \Sigma_1 &= 2t(I + tJv(x)^T + t^2(Jv(x)^T)^2 + O(t^3))(I + tJv(x) + t^2Jv(x)^2 + O(t^3)) \\ &= 2tI + 2t^2(Jv(x)^T + Jv(x)) + 2t^3\left((Jv(x)^T)^2 + Jv(x)^2 + Jv(x)^TJv(x)\right) + O(t^4). \end{aligned}$$

Then we have:

$$\mathrm{Tr}\,(\Sigma_1) = 2Nt - 4t^2\mathrm{Tr}\,(Jv(x)) + 2t^3(2\mathrm{Tr}\left(Jv(x)^2\right) + \mathrm{Tr}\left(Jv(x)^T Jv(x)\right)),$$

and

$$\mathrm{Tr}\,(\Sigma_2) = 2Nt.$$

Furthermore, $\mathrm{Tr}\left([\sqrt{\Sigma_1}\Sigma_2\sqrt{\Sigma_1}]^{1/2}\right) = \sqrt{2t}\mathrm{Tr}\left(\sqrt{\Sigma_1}\right)$, with:

$$\begin{aligned}\sqrt{\Sigma_1} &= \sqrt{2t}(I - \frac{1}{2}tJv(x)^T + \frac{3}{8}t^2(Jv(x)^T)^2O(t^3))(I - \frac{1}{2}tJv(x)^T + \frac{3}{8}t^2Jv(x)^2 + O(t^3)) \\ &= \sqrt{2t}\left(I - \frac{1}{2}t(Jv(x)^T + Jv(x)) + O(t^2)\right).\end{aligned}$$

Then

$$\begin{aligned}\Sigma_2^{1/2} - \Sigma_1^{1/2} &= \sqrt{2t}I - \sqrt{2t}(I + t(Jv(x) + Jv(x)^T) + t^2Jv(x)Jv(x)^T)^{1/2} \\ &= -\frac{\sqrt{2}}{2}t\sqrt{t}(Jv(x) + Jv(x)^T) + o(t\sqrt{t}).\end{aligned}$$

Hence $\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|^2_{\text{Frobenius}} = C_\Sigma t^3 + o(t^3)$ where C_Σ is a constant depending only on v. In a finite dimension, the norms $\mathcal{W}_2$ and $\mathcal{W}_2^{a,b}$ are equivalent, hence

$$\mathcal{W}_2^{a,b}(\mathcal{N}_1, \mathcal{N}_2) \leq C_g\mathcal{W}_2(\mathcal{N}_1, \mathcal{N}_2) \leq \tilde{C}_\Sigma t\sqrt{t}. \tag{2.9}$$

Now plugging in the estimates (2.6), (2.7), (2.8) and (2.9) into equation (2.4) and by equivalence of the norms, we showed that there exists $\beta > 0$ such that

$$\mathcal{W}_2(\mu_1, \mu_2) \leq \beta t\sqrt{t}.$$

□

Another useful estimate is that of the Wasserstein distance between a measure $\mu \in \mathcal{P}(\mathcal{M})$ and its convolution with the heat kernel of $\mathcal{M}$. We have the following result (on a fixed manifold $\mathcal{M}$):

Lemma 2.1.3. *Let $\mathcal{M}$ be a Riemannian manifold, with Ricci curvature globally bounded below. Define its heat kernel by $p^t(x, y)dy$, providing the solution at time t of the heat equation $\partial_\tau\mu = \Delta_\mathcal{M}\mu$ by convolution with the initial data (where $\Delta_\mathcal{M}$ denotes the Laplace-Beltrami operator of $\mathcal{M}$). Let*

$\mu_0 \in \mathcal{P}(\mathcal{M})$. *There exists* $C > 0$ *independent of* $\mathcal{M}$ *such that for* t *small enough,*

$$\mathcal{W}(\mu_0, p^t * \mu_0) \leq C\sqrt{\mathrm{vol}(\mathcal{M})}\sqrt{t}.$$

Proof. First remark that the boundedness from below of the Ricci curvature ensures the existence of a unique heat kernel $p^t(x,y)$ on $\mathcal{M}$ (see [129]). We evaluate the distance between a measure and its convolution with the heat kernel. To estimate it, we remind estimates on the distance between a measure μ on a Riemannian manifold $\mathcal{M}$ and its convolution with the heat kernel $p^t(x,y)dy$ where $\mu(t,x) = \mu_0 * p^t(x,y) = \int_{\mathcal{M}} p^t(x,y)d\mu_0(y)$. From the definition of the Wasserstein distance, $W^2(\mu_0, \mu_0 * p^t) \leq \int_{\mathcal{M}\times\mathcal{M}} d(x,y)^2 d\pi(x,y)$ for all transference plan with marginals μ_0 and $\mu_0 * p^t$. Let $\pi(x,y) = p^t(x,y)d\mu_0(y)dy$. We show easily that its marginals are μ_0 and $\mu_0 * p^t$. Indeed, for all $E \subset \mathcal{M}$,

$$\pi(E \times \mathcal{M}) = \int_{E\times\mathcal{M}} d\mu_0(x)p^t(x,y)dy = \int_E d\mu(0,x) = \mu(0,E)$$

and

$$\pi(\mathcal{M} \times E) = \int_{\mathcal{M}\times E} d\mu(0,x)p^t(x,y)dy = (\mu * P)(t,E).$$

Hence

$$W^2(\mu_0, \mu_0 * p^t) \leq \int_{\mathcal{M}\times\mathcal{M}} d(x,y)^2 d\mu_0(x)p^t(x,y)dy.$$

Varadhan's estimate for the heat kernel on a close manifolf $\mathcal{M}$ (see [79]) gives:

$$\lim_{t\to 0}\left(-2t\ln(p^t(x,y))\right) = d(x,y)^2.$$

Hence for t small enough, $p^t(x,y) \leq e^{-\frac{d(x,y)^2}{4t}}$. Then

$$\begin{aligned}\mathcal{W}^2(\mu_0, \mu_0 * p^t) &\leq \int_{\mathcal{M}} p^t(x,y)d(x,y)^2dy \leq \int_{\mathcal{M}} e^{-d(x,y)^2/(4t)}d(x,y)^2dy \leq \int_{\mathcal{M}} 4e^{-1}tdy \\ &\leq 4e^{-1}\mathrm{vol}(\mathcal{M})\, t.\end{aligned}$$

□

In order to prove Theorem 2.1.1, we first prove the following:

Lemma 2.1.4. *Let* $s \in [0,T]$. *Then there exists* $\mu_s^* := \lim_{n\to\infty} \mu_s^n$. *Moreover,* μ^* *is a continuous curve in* $\mathcal{P}_c(\mathbb{R}^d)$, *satisfying* $\mu_0^* = \mu_0$.

The proof of Lemma 2.1.4 requires the use of the Arzelà-Ascoli theorem and Prokhorov's theorem. We recall the Arzelà-Ascoli theorem generalized for topological vector spaces [62].

Theorem 2.1.2 (Arzelà-Ascoli)**.** *Let X be a compact Hausdorff space and Y a metric space. Then a family $F \subset \mathcal{C}(X, Y)$ is relatively compact in the compact-open topology if and only if:*

1. *F is pointwise relatively compact*
2. *F is equicontinuous*

Let us also recall the definition of tightness.

Definition 2.1.2. *A set of measures P is tight if and only if for all $\epsilon > 0$, there exists a compact K such that for all $\mu \in P$, $\mu(\mathbb{R}^d \setminus K) \leq \epsilon$.*

Lastly, Prokhorov's theorem states:

Theorem 2.1.3 (Prokhorov)**.** *Let X be a Polish space (i.e. a separated, completely metrizable topological space). A set P in the space of probability measures $\mathcal{P}(X)$ is relatively compact if and only if it is tight.*

We now prove Lemma 2.1.4.

Proof. We will use the Arzelà-Ascoli theorem to prove that there exists a converging subsequence of (μ^n). In our setting, $X = [0, T]$ is a compact Hausdorff space, and we take $Y = \mathcal{P}(\mathbb{R}^d)$, the space of probability in $\mathbb{R}^d$, endowed with the Wasserstein metric $\mathcal{W}_p$. We study the family $F := (\mu^n)_{n\in\mathbb{N}} \in \mathcal{C}([0, T], \mathcal{P}(\mathbb{R}^d))$.

(i) We start by showing that F is pointwise relatively compact.
Let $t \in [0, T]$. According to Prokhorov's theorem, $F \subset \mathcal{P}(\mathbb{R}^d)$ is relatively compact if and only if it is tight. Each $\mu^n(t)$ is compactly supported. Hence, to prove that $(\mu^n(t))_{n\in\mathbb{N}}$ is tight, we need to show that $\mathcal{M}_t^n := \text{supp}(\mu^n(t))$ is bounded independently of n. Let $n \in \mathbb{N}$ and $x_t^n \in \mathcal{M}_t^n$. Let $s \in [0, t_n)$ and $l \in \{0, ..., 2^n - 1\}$ such that $t = lt_n + s$. Then there exists $x_0 \in \mathcal{M}_0$ such that $x_t^n = \phi_s^{n,l}(\phi_{t_n}^{n,l-1}(\phi_{t_n}^{n,l-2}(...\phi_{t_n}^{n,0}(x_0))...)))$. Since the vector field $v[\cdot]$ is bounded, we have $\|x_t^n\| \leq \|x_0\| + \|v[\cdot]\|_\infty t$. Hence $\mathcal{M}_t^n$ is bounded independently of n. This implies that for $t \in [0, T]$, $(\mu^n(t))_{n\in\mathbb{N}}$ is tight. So $(\mu^n)_{n\in\mathbb{N}}$ is pointwise relatively compact.

(ii) Let us now show that F is equicontinuous (equivalent to uniform equicontinuity since $[0,T]$ is compact). We will prove that each μ^n is 1/2-Hölder as function of time with values in the space of probability measures (endowed with the Wasserstein distance).

We give several estimates that will prove useful in what follows.

Let $\mu, \nu \in \mathcal{P}_c(\mathbb{R}^d)$. The distance between a measure and its transport was was estimated in [95]:

$$\mathcal{W}(\mu, \phi_v^t \# \mu) \leq \|v\|_\infty t, \tag{2.10}$$

where we use the boundedness of v. Furthermore

$$\mathcal{W}(\phi_{v[\mu]}^t \# \mu, \phi_{v[\nu]}^t \# \nu) \leq \left(e^{\frac{3}{2}Lt} + \frac{K}{L} e^{\frac{L}{2}t}(e^{Lt} - 1) \right) \mathcal{W}(\mu, \nu). \tag{2.11}$$

From Lemma 2.1.3, $\mathcal{W}(\mu, \mu * p^t) \leq C\sqrt{\mathrm{vol}(\mathcal{M})}\sqrt{t}$. Here, since v is uniformly bounded, there exists $\bar{C} > 0$ independent of $\mathcal{M}$ such that

$$\mathcal{W}(\mu, \mu * p^t) \leq \bar{C}\sqrt{t}. \tag{2.12}$$

Now we give a similar estimate as (2.11) but for the heat evolution. From [38], for $\mu, \nu \in \mathcal{P}(\mathcal{M})$ where $\mathcal{M}$ is a smooth, connected and complete Riemannian manifold with Ricci curvature bounded from below, i.e. $\mathrm{Ric}(\mathcal{M}) \geq \kappa$, we have:

$$\mathcal{W}(p^t * \mu, p^t * \nu) \leq e^{\kappa t}\mathcal{W}(\mu, \nu). \tag{2.13}$$

Let $n \in \mathbb{N}$ and $t_n = 2^{-n}T$. Let $l \in \mathbb{N}$ and $k \in \mathbb{N}$ such that $(l+k)t_n \leq T$. We want to estimate the Wasserstein distance between $\mu^n(lt_n)$ and $\mu^n((lt_n + s)$. Suppose that $s = 2^{k-1}t_n$ for some $k \in \mathbb{N}$. Let $\mu^n(lt_n)$ be fixed. For economy of notation, we will also denote it by μ_l^n. Then there exists 2^{k-1} flows and 2^{k-1} heat kernels such that

$$\mathcal{W}_2(\mu(lt_n), \mu((l+m)t_n)) = \mathcal{W}_2(\mu(lt_n), \phi_{2^k}^{t_n} \# p_{2^k-1}^{t_n} * ... * \phi_2^{t_n} \# p_1^{t_n} * \mu(lt_n)).$$

We prove by induction on k that for each k there exists R_k such that

$$\mathcal{W}_2(\mu(lt_n), \phi_{2^k}^{\tau} \# p_{2^k-1}^{\tau} * ... * \phi_2^{\tau} \# p_1^{\tau} * \mu(lt_n)) \leq R_k(s + \sqrt{s}) \tag{2.14}$$

where $\tau = 2^{-k+1}s$.

For $k = 1$, let $\tau = s$. From (2.10) and (2.12), we have:

$$\begin{aligned}\mathcal{W}_2(\mu(lt_n), \phi_2^\tau \# p_1^\tau * \mu(lt_n)) &\leq \mathcal{W}_2(\mu(lt_n), p_1^\tau * \mu(lt_n)) + \mathcal{W}_2(p_1^\tau * \mu(lt_n), \phi_2^\tau \# p_1^\tau * \mu(lt_n)) \\ &\leq \bar{C}\sqrt{\tau} + \|v\|_\infty \tau.\end{aligned}$$

and we set $R_1 := \max(\bar{C}, \|v\|_\infty)$.

Now suppose that (2.14) holds true for some $k \in \mathbb{N}$. Let $\tau = 2^{-k}s$. Then

$$\begin{aligned}&\mathcal{W}_2(\mu_l^n, \phi_{2^{k+1}}^\tau \# p_{2^{k+1}-1}^\tau * ... * \phi_2^\tau \# p_1^\tau * \mu_l^n) \\ =&\mathcal{W}_2\left(\mu_l^n, \prod_{j=0}^{2^{k+1}-1} (\phi_{2^{k+1}-4j}^\tau p_{2^{k+1}-4j-1}^\tau \phi_{2^{k+1}-4j-2}^\tau p_{2^{k+1}-4j-3}^\tau)\mu_l^n\right) \\ \leq&\mathcal{W}_2\left(\mu_l^n, \prod_{j=0}^{2^{k-1}-1} (\phi_{2^{k+1}-4j}^\tau \phi_{2^{k+1}-4j-2}^\tau p_{2^{k+1}-4j-3}^{2\tau})\mu_l^n\right) + \\ &\mathcal{W}_2\Bigg(\prod_{j=0}^{2^{k-1}-1} (\phi_{2^{k+1}-4j}^\tau \phi_{2^{k+1}-4j-2}^\tau p_{2^{k+1}-4j-3}^{2\tau})\mu_l^n, \\ &\prod_{j=0}^{2^{k-1}-1} (\phi_{2^{k+1}-4j}^\tau p_{2^{k+1}-4j-1}^\tau \phi_{2^{k+1}-4j-2}^\tau p_{2^{k+1}-4j-3}^\tau)\mu_l^n\Bigg).\end{aligned} \tag{2.15}$$

The first term of (2.15) can be estimated using the induction hypothesis for k:

$$A_1 = \mathcal{W}_2\left(\mu_l^n, \prod_{j=0}^{2^{k-1}-1} (\phi_{2^{k+1}-4j}^\tau \phi_{2^{k+1}-4j-2}^\tau p_{2^{k+1}-4j-3}^{2\tau})\mu_l^n\right) \leq R_k(s + \sqrt{s}).$$

For the second term of (2.15) we write:

$$\begin{aligned}A_2 = \mathcal{W}_2\Bigg(&\prod_{j=0}^{2^{k-1}-1} (\phi_{2^{k+1}-4j}^\tau \phi_{2^{k+1}-4j-2}^\tau p_{2^{k+1}-4j-3}^{2\tau})\mu_l^n, \\ &\prod_{j=0}^{2^{k-1}-1} (\phi_{2^{k+1}-4j}^\tau p_{2^{k+1}-4j-1}^\tau \phi_{2^{k+1}-4j-2}^\tau p_{2^{k+1}-4j-3}^\tau)\mu_l^n\Bigg) \\ \leq \sum_{i=1}^{2^{k-1}-1} \mathcal{W}_2\Bigg(&P_i^- \phi_{2^{k+1}-4i}^\tau \phi_{2^{k+1}-4i-2}^\tau p_{2^{k+1}-4i-3}^{2\tau} P_i^+ \mu_l^n, \\ &P_i^- \phi_{2^{k+1}-4i}^\tau p_{2^{k+1}-4i-1}^\tau \phi_{2^{k+1}-4i-2}^\tau p_{2^{k+1}-4i-3}^\tau P_i^+ \mu_l^n\Bigg)\end{aligned}$$

where P_i^- and P_i^+ are two operators defined as:

$$\begin{cases} P_i^- = \prod_{j=0}^{i-1}(\phi^\tau_{2^{k+1}-4j}p^\tau_{2^{k+1}-4j-1}\phi^\tau_{2^{k+1}-4j-2}p^\tau_{2^{k+1}-4j-3}) \\ P_i^+ = \prod_{j=i+1}^{2^{k-1}-1}(\phi^\tau_{2^{k+1}-4j}p^\tau_{2^{k+1}-4j-1}\phi^\tau_{2^{k+1}-4j-2}p^\tau_{2^{k+1}-4j-3}). \end{cases}$$

From (2.11) and (2.13), we write:

$$\begin{aligned} &\mathcal{W}_2\Big(P_i^-\phi^\tau_{2^{k+1}-4i}\phi^\tau_{2^{k+1}-4i-2}p^{2\tau}_{2^{k+1}-4i-3}P_i^+\mu_l^n, \\ &\qquad\qquad P_i^-\phi^\tau_{2^{k+1}-4i}p^\tau_{2^{k+1}-4i-1}\phi^\tau_{2^{k+1}-4i-2}p^\tau_{2^{k+1}-4i-3}P_i^+\mu_l^n\Big) \\ &\le e^{2i(L-\kappa)\tau+L\tau}\mathcal{W}_2\left(\phi^\tau_{2^{k+1}-4i-2}p^\tau_{2^{k+1}-4i-3}\tilde{\mu}^i(lt_n), p^\tau_{2^{k+1}-4i-1}\phi^\tau_{2^{k+1}-4i-2}\tilde{\mu}^i(lt_n))\right) \\ &\le e^{2i(L-\kappa)\tau+L\tau}\beta\tau\sqrt{\tau} \end{aligned}$$

where $\tilde{\mu}^i(lt_n) := p^\tau_{2^{k+1}-4i-3}P_i^+\mu(lt_n)$ and the last inequality comes from the bracket estimation in Lemma 2.1.2. Then

$$\begin{aligned} A_2 \le &\sum_{i=1}^{2^{k-1}-1} e^{2i(L-\kappa)\tau+L\tau}\beta\tau\sqrt{\tau} = e^{L\tau}\beta\tau\sqrt{\tau}\sum_{i=1}^{2^{k-1}-1} e^{2i(L-\kappa)\tau} = e^{L\tau}\beta\tau\sqrt{\tau}\frac{e^{2(L-\kappa)\tau(2^{k-1}-1)}-1}{e^{2(L-\kappa)\tau}-1} \\ \le &e^{L\tau}\beta\tau\sqrt{\tau}\frac{e^{2(L-\kappa)\tau(2^{k-1}-1)}}{2(L-\kappa)\tau} = e^{L\tau}\beta\frac{1}{2(L-\kappa)}\sqrt{\tau}e^{(L-\kappa)s-2\tau(L-\kappa)} = \frac{e^{(-L+2\kappa)\tau}\beta}{2(L-\kappa)}\sqrt{\tau}e^{(L-\kappa)s} \\ \le &J\sqrt{\tau} \end{aligned}$$

where $J = \frac{\max(1, e^{(-L+2\kappa)s})\beta}{2(L-\kappa)}e^{(L-\kappa)s}$. Plugging in A_1 and A_2 in (2.15), we have:

$$\begin{aligned} &\mathcal{W}_2(\mu(lt_n), \phi^\tau_{2^{k+1}}\#p^\tau_{2^{k+1}-1} * ... * \phi^\tau_2\#p^\tau_1 * \mu(lt_n)) \le A_1 + A_2 \le R_k(s+\sqrt{s}) + J\sqrt{2^{-k}s} \\ &\le (R_k + 2^{-k/2}J)(s+\sqrt{s}) = R_{k+1}(s+\sqrt{s}). \end{aligned}$$

Hence (2.14) is satisfied for all $k \in \mathbb{N}$. Furthermore, notice that $(R_k)_{k\in\mathbb{N}}$ is a converging sequence:

$$R_k = \sum_{i=2}^{k} 2^{-i/2}J + R_1 \xrightarrow[k\to\infty]{} \bar{R} := \frac{J}{1-2^{-1/2}} + R_0$$

Since (R_k) is monotonically increasing, $R_k \le \bar{R}$ for all $k \in \mathbb{N}$. Hence for $s = 2^{k-1}t_n$,

$$\mathcal{W}_2(\mu(lt_n), \mu(lt_n+s)) = \mathcal{W}_2(\mu(lt_n), \phi^{t_n}_{2^{k+1}}\#p^{t_n}_{2^{k+1}-1} * ... * \phi^{t_n}_2\#p^{t_n}_1 * \mu(lt_n)) \le \bar{R}(s+\sqrt{s})$$

where $\bar{R}$ is independent of n, l and s. This proves that each μ_n is Hölder of order 1/2. Thus F

is equicontinuous and according to the Arzelà-Ascoli theorem, there exists a limit to the sequence $(\mu_n)_{n\in\mathbb{N}}$.

□

We now state the main result of this section, that describes locally the dynamics of μ^*.

Theorem 2.1.4. *Let μ^* be defined in Lemma 2.1.4, with a given initial data $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$. Then we have*

$$\lim_{s\to 0} \frac{\mu_s^* - e^{\Delta_0 s}\mu_0}{s} = -\nabla\cdot(v\mu_s). \tag{2.16}$$

Hence μ^* is the unique solution of

$$\begin{cases} \partial_s \mu_s + \nabla\cdot(v\mu_s) = \Delta_s \mu_s \\ \mu(s=0) = \mu_0. \end{cases} \tag{2.17}$$

2.1.2 The intrinsic Laplace-Beltrami operator

We describe the evolution of the cell as follows: consider its shape at time 0 and call it $M = M_0$, that is an oriented compact manifold, and a sub-manifold of $\mathbb{R}^d$. Its shape then evolves by the flow Φ_t of a vector field v defined on the whole $\mathbb{R}^d$, or at least in a whole neighborhood of M_0 as a subset of $\mathbb{R}^n$. The new shape is then $M_t = \Phi_t(M_0)$, that is a sub-manifold of $\mathbb{R}^n$ since Φ_t is a diffeomorphism.

Observe that both M_0 and M_t are Riemannian manifold, with Riemannian structure induced by their embedding in $\mathbb{R}^n$. We then represent the structure as follows: we fix the manifold $M = M_0$ and change its Riemannian structure $< .,. >_t$ with respect to time. We have

$$\Phi_t : \begin{cases} M & \to & \mathbb{R}^n \\ x & \mapsto & \Phi_t(x) \end{cases}$$

and we endow M with the Riemannian structure $< .,. >_t$ induced by the pull-back of the Riemannian structure of $\Phi_t(M) = M_t \subset \mathbb{R}^n$. This implies

$$< w_1, w_2 >_t = < \Phi_t^* w_1, \Phi_t^* w_2 >_E, \tag{2.18}$$

where $< .,. >_E$ is the standard Euclidean structure in $\mathbb{R}^n$ (or eventually any other Riemannian metric in $\mathbb{R}^n$).

The manifold $(M, <.,.>_t)$ is Riemannian, then the Laplace-Beltrami operator Δ_t is intrinsically defined. We are interested in describing such operator as a function of t. Recalling that it is the divergence of the gradient, we aim at describing div^t and grad^t as a function of time. In particular, we aim at computing first-order development of such operators with respect to time.

Remark 2.1.2. *The operator Δ_t is intrinsically defined as the divergence of the gradient for the time-evolving metric g_t on the fixed manifold $\mathcal{M}_0$. It differs from the Laplace-Beltrami operator used in Chapter 1, which is computed as the divergence of the gradient on the time-evolving manifold $\mathcal{M}_t$, for the metric inherited from the embedding in the ambient Euclidean space. In a parallel line of work not presented in this thesis, we compare these two intrinsic and extrinsic operators.*

We first have

$$\Phi_t^* w = w + tJv \cdot w + o(t),$$

where J is the Jacobian with respect to the Euclidean structure of $\mathbb{R}^n$ and $\cdot$ represents the linear action of the linear operator Jv on w. This implies

$$< w_1, w_2 >_t = < w_1, w_2 >_E + t\,(< Jv \cdot w_1, w_2 >_E + < Jv \cdot w_2, w_1 >_E) + o(t).$$

Since vectors w_1, w_2 belong to T_xM, we will denote with $J_M v$ the restriction of Jv to T_xM by projection, i.e.

$$J_M v : \begin{cases} T_xM & \to & T_xM \\ w & \mapsto & (Jv \cdot w)_M, \end{cases}$$

where z_M is the component of the vector $z \in T_x\mathbb{R}^n$ on the subspace T_xM. Observe that we are using here the Riemannian structure of $\mathbb{R}^n$ to define projections.

We now study the gradient $\mathrm{grad}^t f$ for a function $f \in C^\infty(M)$, via its intrinsic definition. For all $w \in T_xM$ it holds

$$< \mathrm{grad}^t(f), w >_t = \mathcal{L}_w f$$

Since the identity at time $t = 0$ holds for grad^0, we have $\mathrm{grad}^t f = \mathrm{grad}^0 f + tB_1$, for a vector field

B_1 to be found, We have in particular

$$< \mathrm{grad}^0(f) + tB_1 + tJv \cdot \mathrm{grad}^0(f) + o(t), w + t\, Jv \cdot w + o(t) >_E = \mathcal{L}_w f$$
$$t\left(< B_1, w >_E + < Jv \cdot \mathrm{grad}^0(f), w >_E + < \mathrm{grad}^0 f, Jv \cdot w >_E\right) + o(t) = 0.$$

We then have $B_1 = -(Jv \cdot \mathrm{grad}^0(f))_M - B$ where $(Jv \cdot \mathrm{grad}^0(f))_M$ is the component of $Jv \cdot \mathrm{grad}^0(f)$ on the tangent space of M, and $B(f, v)$ is intrinsically defined by the following rule: for all $w \in T_x M$ it holds

$$< B(f, v), w >_E = < \mathrm{grad}^0(f), Jv \cdot w >_E . \tag{2.19}$$

Summing up, we have

$$\mathrm{grad}^t(f) = \mathrm{grad}^0(f) - t(Jv \cdot \mathrm{grad}^0(f))_M - tB(f, v) + o(t) \tag{2.20}$$

with $B(f, v)$ defined by (2.19).

We now study the divergence $\mathrm{div}^t(X)$ for a vector field $X \in \mathrm{Vec}\,(M)$. Given vol_t the volume form of the Riemannian manifold, it holds

$$\mathrm{div}^t(X)\mathrm{vol}_t = \mathcal{L}_X \mathrm{vol}_t. \tag{2.21}$$

Observe that $\mathrm{vol}_t = \sqrt{|g^t|}\, dX_1 \wedge dX_2 \wedge \ldots \wedge dX_m$ for any base $X_1, \ldots, X_m$ of the Riemannian manifold $(M, < ., . >_t)$. We choose an orthonormal basis for $(M, < ., . >_0)$ and study the evolution of vol_t. Since $g^0 = \mathrm{Id}$ and $|\mathrm{Id} + tA| = 1 + t\mathrm{Tr}\,(A) + o(t)$, we have

$$|g^t| = 1 + 2t \sum_{i=1}^{m} < Jv \cdot X_i, X_i >_E = 1 + 2t\mathrm{Tr}\,(Jv)_M ,$$

where the operator Jv is restricted to the tangent space of M. This implies $\sqrt{|g^t|} = 1 + t\mathrm{Tr}\,(Jv)_M$. Writing $\mathrm{div}^t(X) = \mathrm{div}^0 X + tf + o(t)$ for a function f to be found, it holds from (2.21)

$$\begin{aligned}((\mathrm{div}^0(X) + tf)(1 + t\mathrm{Tr}\,(Jv)_M) + o(t))\mathrm{vol}_0 &= (\mathcal{L}_X(1 + t\mathrm{Tr}\,(Jv)_M))\,\mathrm{vol}_0 + (1 + t\mathrm{Tr}\,(Jv)_M)\mathcal{L}_X \mathrm{vol}_0 \\ &= (\mathcal{L}_X(1 + t\mathrm{Tr}\,(Jv)_M))\,\mathrm{vol}_0 + (1 + t\mathrm{Tr}\,(Jv)_M)\mathrm{div}^0(X)\mathrm{vol}_0,\end{aligned}$$

hence

$$\mathrm{div}^t(X) = \mathrm{div}^0(X) + t\mathcal{L}_X \mathrm{Tr}\,(Jv)_M\,. \tag{2.22}$$

Observe that this formula is intrinsic, since the trace of the linear operator Jv does not depend on the chosen orthonormal frame.

We now compute the Laplace-Beltrami operator Δ^t. Since $\Delta^t f = \mathrm{div}^t(\mathrm{grad}^t(f))$ by definition, and observing that it holds $L_{\mathrm{grad}^0(f)}\mathrm{Tr}\,(Jv)_M =< \mathrm{grad}^0(f), \mathrm{grad}^0(\mathrm{Tr}\,(Jv)_M) >_E$, we have

$$\Delta^t(f) = \Delta^0(f) + t\left(< \mathrm{grad}^0(f), \mathrm{grad}^0(\mathrm{Tr}\,(Jv)_M) >_E - \mathrm{div}^0(B(f,v) + (Jv \cdot \mathrm{grad}^0(f))_M)\right) + o(t).$$

2.2 Noncommutativity of heat and transport evolutions

2.2.1 Commutator of the Laplace-Beltrami operator with the vector field

In this section we consider the following phenomenon. Let us assume that one is allowed to vary slightly the values of t, τ (that are no more identical) in the scheme $\mathbb{S}$. Then, we will prove then the final value of the scheme $\mathbb{S}$ will be different than the (unique) of (2.17). In terms of reachability: authorising a slight variations of t, τ implies a larger set of attainable final configurations. The goal of this section is to prove that the set of attainable configurations is larger, and to describe such set.

For finite dimensional systems, this phenomenon is studied for the so-called switching systems. Consider two vector fields X_0, X_1 and a measurable switching function $u : [0,T] \to \{0,1\}$. Then one can consider the dynamics of the system

$$\begin{cases} \dot{x} = X_{u(t)} \\ x(0) = x_0. \end{cases}$$

The solution at time T of this system, denoted by $x(T,u)$, is unique. Nevertheless, if one chooses another switching function $v : [0,T] \to \{0,1\}$, one has in general $x(T,u) \neq x(T,v)$. Then, it is of interest to study the set of all possible final states, at least for small T. A classical result in control theory, the Orbit theorem, (roughly) states that the set of attainable configurations is related with the Lie bracket $[X_0, X_1]$, and in particular that one can choose good switching functions to drive the system along a direction arbitrarily close to the vector field $[X_0, X_1]$.

For this reason, we study in this thesis the bracket between the "heat vector field" and the

"transport vector field". Indeed, one can consider the solution of an heat equation as a continuous (and even differentiable) curve in $P_2(X)$ endowed with the Wasserstein distance. The, the time derivative of this curve in a point μ_t (that is clearly the Laplacian $\Delta\mu_t$) can be considered as a vector field, that we call the **heat vector field**. Similarly, we define the **transport vector field** as the derivative of the solution of the transport equation in a point.

By borrowing the notation from Lie brackets of vector fields, we define

$$[\Delta, v]\mu := \lim_{t,\tau\to 0} \frac{\Phi_{-t}\#\left(e^{\tau\Delta^t}(\Phi_t\#\mu)\right) - e^{\tau\Delta^0}\mu}{t\tau}, \tag{2.23}$$

where $\Phi_t\#$ is the push-forward of a measure via the flow generated by the vector field v, and $e^{\tau\Delta^t}$ is the semigroup generated by Δ^t at time τ.

Observe that, for any measure $\mu \in \mathcal{M}_0(\mathbb{R}^{2d})(M_0)$, it holds $\Phi_t\#\mu \in \mathcal{M}_0(\mathbb{R}^{2d})(M_t)$. If one chooses the coordinates on the manifold M_t induced by the diffeomorphism Φ_t, then it holds $\Phi_t\#\mu = \mu$ in the sense that their expression with respect to coordinates is the same. Similarly, it holds $\Phi_{-t}\#\mu = \mu$ for any measure $\mu \in \mathcal{M}_0(\mathbb{R}^{2d})(M_t)$.

Then, for any test function $f \in C_c^\infty(M)$, one can write (2.23) as follows:

$$\begin{aligned}
([\Delta, v]\mu)(f) &= \lim_{t,\tau\to 0}\int_M \frac{f}{t\tau}\, d\left(\Phi_{-t}\#\left(e^{\tau\Delta^t}(\Phi_t\#\mu)\right) - e^{\tau\Delta^0}\mu\right) = \lim_{t,\tau\to 0}\int_M \frac{f}{t\tau}\, d\left(e^{\tau\Delta^t}\mu - e^{\tau\Delta^0}\mu\right) \\
&= \lim_{t\to 0}\int_M \frac{f}{t}\, d\left(\Delta^t\mu - \Delta^0\mu\right) = \lim_{t\to 0}\int_M (\Delta^t - \Delta^0)\frac{f}{t}\, d\mu \\
&= \int_M < \mathrm{grad}^0(f), \mathrm{grad}^0(\mathrm{Tr}\,(Jv)_M) >_E - \mathrm{div}^0(B(f,v) + (Jv\cdot \mathrm{grad}^0(f))_M)\, d\mu, \qquad (2.24)
\end{aligned}$$

where we used $\int f\, d\Delta\mu = \int \Delta f\, d\mu$ as the definition of the Laplace-Beltrami operator as an operator on the space of measures. Then (2.24) is the intrinsic formula for the bracket (2.23).

2.2.2 An example: the sphere S^1 in $\mathbb{R}^2$

We now compute the bracket $[\Delta, v]$ for an example. We consider the unit circle S^1 in $\mathbb{R}^2$ parametrized by an angle θ as the initial manifold M, and the vector field $v = (x-1, 2y)$. It is easy to verify that at time t the unit circle is transported to an ellipse of equation: $\left(\frac{x-x_c}{e^t}\right)^2 + \left(\frac{y}{e^{2t}}\right)^2 = 1$ where $x_c = 1 - e^t$ (see Fig. 2.1).

First, we fix the constant initial data $\mu_0 = d\theta$ as the Riemannian volume form on the sphere.

We consider the Euclidean metric on $\mathbb{R}^2$, i.e. Riemannian structure given by the orthonormal frame ∂_x, ∂_y at each point. The corresponding Riemannian structure on S^1 is given by $\partial_\theta = -y\partial_x +$

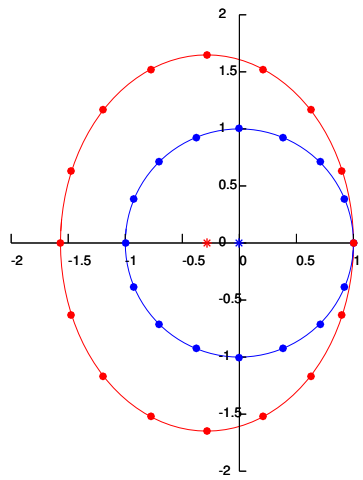


Figure 2.1: Transport of the unit sphere (in blue) by the vector field $v(x,y) := (x-1, 2y)$. At $t = 0.25$, the resulting ellipse (red) is centered at $(1 - e^{0.25}, 0)$.

$x\partial_y$. This implies

$$(J_M v) \cdot \delta_\theta = \begin{pmatrix} -y \\ 2x \end{pmatrix}_M = (1 + x^2)\partial_\theta,$$

thus $\mathrm{Tr}\,(J_M v) = 1 + \cos^2$. Since the initial data is the Riemannian volume form, the divergence theorem implies

$$\begin{aligned} ([\Delta, v]\mu)(f) &= \int_M < \frac{\partial f}{\partial \theta}\partial_\theta, \frac{\partial(1+\cos^2(\theta))}{\partial\theta}\partial_\theta >_E + 0\, d\theta \\ &= \int_M -f\frac{\partial^2 \cos^2(\theta)}{\partial\theta^2}\, d\theta = \int_M 2f\cos(2\theta)\, d\theta. \end{aligned}$$

Hence for an initially constant signal, $[\Delta, v] = 2\cos(2\theta)$.

Now for a more complicated initial data $\mu_0 = (1 + \cos(\theta))d\theta$, the second term in (2.24) is no longer 0. First, notice that:

$$< \mathrm{grad}^0(f), \mathrm{grad}^0(\mathrm{Tr}\,(Jv)_M) >=< \frac{\partial f}{\partial\theta}\partial_\theta, \frac{\partial(1+\cos^2(\theta))}{\partial\theta}\partial_\theta >= \frac{\partial f}{\partial\theta}(-2\cos\theta\sin\theta).$$

Secondly, we calculate the term $B(f,v)$ knowing that $\langle B(f,v), w)\rangle_E = \langle \mathrm{grad}^0(f), Jv \cdot w\rangle_E$. Taking

a vector $w = (w_x, w_y)^T$, we write:

$$B_x w_x + B_y w_y = \langle \mathrm{grad}^0(f), Jv \cdot w \rangle_E = \langle \begin{pmatrix} \partial_x f \\ \partial_y f \end{pmatrix}, \begin{pmatrix} w_x \\ 2w_y \end{pmatrix} \rangle_E = \partial_x f w_x + 2\partial_y f w_y.$$

Hence we have :

$$B_\theta = \langle \begin{pmatrix} B_x \\ B_y \end{pmatrix}, \partial_\theta \rangle = \langle \begin{pmatrix} \partial_x f \\ 2\partial_y f \end{pmatrix}, \begin{pmatrix} -y \\ xf \end{pmatrix} \rangle = -y\partial_x f + 2x\partial_y f = \sin^2\theta \partial_\theta f + 2\cos^2\theta \partial f = (1+\cos^2\theta)\partial_\theta f.$$

Similarly,

$$(Jv \cdot \mathrm{grad}_0(f))_M = \langle \begin{pmatrix} \partial_x f \\ 2\partial_y f \end{pmatrix}, \partial_\theta \rangle \partial_\theta = (1+\cos^2\theta)\partial_\theta f \; \partial_\theta.$$

So

$$\mathrm{div}^0(B(f,v) + Jv \cdot \mathrm{grad}^0(f))_M = 2\partial_\theta((1+\cos^2\theta)\partial_\theta f) = 2(1+\cos^2\theta)\partial_\theta^2 f - 4\cos\theta\sin\theta\partial_\theta f.$$

We calculate separately the two terms in Equation (2.24) :

$$\begin{aligned}
(i) \int_M &< \mathrm{grad}^0(f), \mathrm{grad}^0(\mathrm{Tr}\,(Jv)_M) > (1+\cos\theta)d\theta = \int_M \frac{\partial f}{\partial\theta}(-2\cos\theta\sin\theta)(1+\cos\theta)d\theta \\
&= \int_M 2f\frac{\partial}{\partial\theta}((\cos\theta\sin\theta)(1+\cos\theta))d\theta = \int_M f(6\cos^3\theta + 4\cos^2\theta - 4\cos\theta - 2)d\theta. \\
(ii) \int_M &-\mathrm{div}^0((B(f,v) + Jv \cdot \mathrm{grad}^0(f))_M)(1+\cos\theta)d\theta \\
&= -2\int_M ((1+\cos^2\theta)\frac{\partial^2 f}{\partial\theta^2} - 2\cos\theta\sin\theta\frac{\partial f}{\partial\theta})\cos\theta d\theta \\
&= 2\int_M [\frac{\partial f}{\partial\theta}(-\sin\theta - 3\cos^2\theta\sin\theta) - 2f(-2\cos\theta\sin^2\theta + \cos^3\theta)]d\theta \\
&= 2\int_M f(-\cos\theta + 3\cos^3\theta)d\theta.
\end{aligned} \tag{2.25}$$

Summing the two terms we get:

$$([\Delta, v]\mu)(f) = \int_M f(12\cos^3\theta + 4\cos^2\theta - 6\cos\theta - 2)d\theta$$

In order to study the bracket $[v, \Delta]$, we use two schemes $\mathbb{S}$ and $\tilde{\mathbb{S}}$ that discretize the diffusion-growth problem described above. We define $\tilde{\mathbb{S}}$ similarly to $\mathbb{S}$ (defined in Sec. 2.1.1, but inverting steps 1 and 2. Hence $\mathbb{S}$ does a series of growth and diffusion operations on the function μ_0 starting

with growth, while $\tilde{\mathbb{S}}$ does the same starting with diffusion. Figure 2.2 shows the first two iterations of each scheme, starting from the same function μ_0 (renamed x_0 and y_0 for notation convenience), and denoting respectively by x_n and y_n the solutions after each iteration of $\mathbb{S}$ and $\tilde{\mathbb{S}}$.

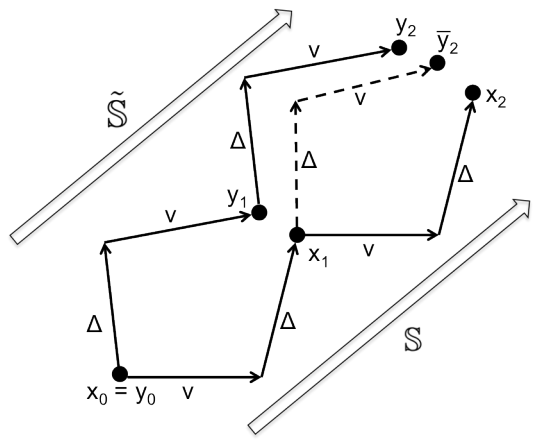


Figure 2.2: Two iterations of the schemes $\mathbb{S}$ and $\tilde{\mathbb{S}}$ starting from the same point $x_0 = y_0$.

We apply this scheme to the signal initially given by the constant function $\mu_0(\theta) = 0.1$.

Numerically, we apply the schemes $\mathbb{S}$ and $\tilde{\mathbb{S}}$ to μ_0 and compute the numerical bracket given by $[\Delta, v]_{\text{num}} = \lim_{\epsilon \to 0}(y_1 - x_1)/\epsilon^2$ (where x_1 and y_1 respectively correspond to the first iterations of $\mathbb{S}$ and $\tilde{\mathbb{S}}$).

Figure 2.3 (left) shows the evolution of the initially constant signal $\mu_0 = 0.1$ after one iteration of each scheme $\mathbb{S}$ and $\tilde{\mathbb{S}}$, with a vector field $v = (x - 1, 2y)$ and time-steps $t = \tau = 0.05$ (so $T = t + \tau = 0.1$).

Figure 2.3 (right) shows the convergence of the bracket when the time step $T = t + \tau$ tends to 0. In the case of the previously defined initially constant signal, the bracket converges to the theoretical value $[\Delta, v]_{\text{theo}}(\mu_0) = 0.2\cos(\theta)$. Figure 2.4 shows the convergence of the bracket for the inital signal $\mu_0(\theta) = 0.1(\cos(\theta)+1)d\theta$. The bracket converges to the theoretical value $[\Delta, v]_{\text{theo}}(\mu_0) = 12\cos^3\theta + 4\cos^2\theta - 6\cos\theta - 2$.

2.3 Control of growth via a signal

2.3.1 Model

The development of an organism is caused by morphogens, signaling molecules that diffuse in the organism and act on cells to produce local responses depending on their local concentration. Growth

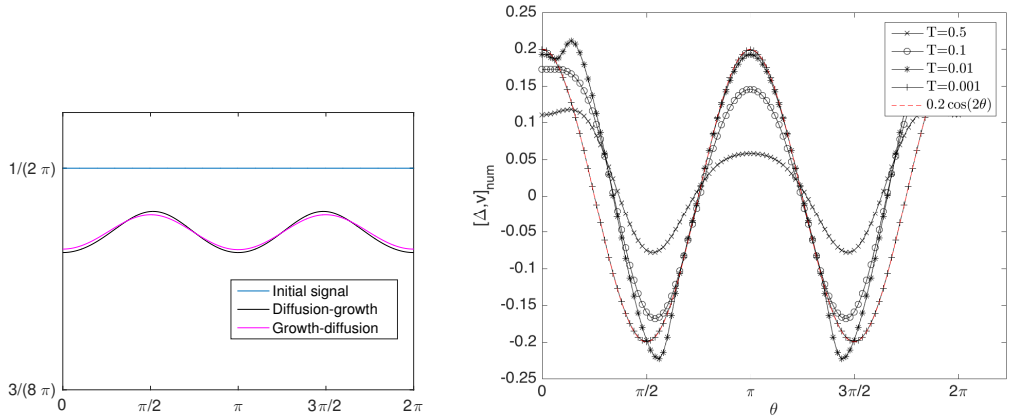


Figure 2.3: Left: Evolution of the signal $\mu_0(\theta) = 0.1$ after a time-step $T = 0.1$ for the two schemes. Right: Convergence of the bracket to the theoretical one for the initial signal $\mu_0(\theta) = 0.1d\theta$.

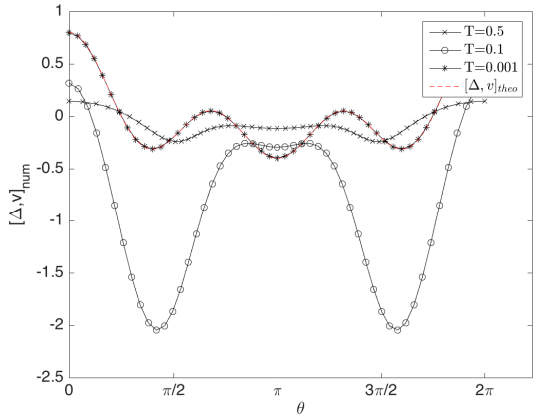


Figure 2.4: Convergence of the bracket for the initial signal $\mu_0(\theta) = 0.1(\cos(\theta) + 1)d\theta$.

is thus induced by the distribution of a signal, and the diffusion of the signal is itself affected by the changing shape and size of the organism. In other words, there is a complete coupling between a PDE describing the signal's evolution and a time-varying manifold. In this section we give a setting to describe this coupling. More specifically, we model the signal's diffusion from a source using the time-dependent Laplace-Beltrami operator defined by the Riemannian structure on the manifold, and the evolution of the manifold depends on the concentration of the signal. Our first objective is to study the controllability of the manifold's shape from a source with variable intensity.

We consider a 1D model to describe a cell membrane. Given an angle variable $\theta \in S^1$, we describe the position of the membrane by a function $r = r(t, \theta)$ representing the radius. We also consider a signal $s = s(t, \theta)$ on the cell, that pushes the cell to grow in its radial direction. Then $\partial_t r = s$.

The dynamics of s are given by the heat equation on the cell. Since the shape of the cell is defined by r, we denote the Laplacian on the cell by Δ_r, the Laplace-Beltrami operator on the cell with shape r. Moreover, our control is the value of s in a given point, say in $\theta = 0$, the point in which the nucleus sends the growing signal to the boundary.

Hence, the dynamics satisfies:

$$\begin{cases} \partial_t r = s, \\ \partial_t s = \Delta_r s, \\ s(t, \theta = \pi) = u(t). \end{cases} \tag{2.26}$$

We now assume that the initial configurations of both r and s are symmetric with respect to θ, i.e. $r(0, -\theta) = r(0, \theta)$ and, similarly, $s(0, -\theta) = s(0, \theta)$. The simplest example is $r(0, \theta) = 1$ and $s(0, \theta) = 0$, i.e. a round cell and a zero signal on it. One can easily prove by that, for any choice of the control $u(t)$, both r and s stay symmetric. Indeed, using the explicit expression of the Laplace-Beltrami operator (2.30), we prove that $(s(t, \theta), r(t, \theta))$ and $((s(t, -\theta), r(t, -\theta))$ solve the same differential system. Since the two couples have identical initial conditions, by uniqueness of solution we deduce that they are equal and thus symmetric.

Since s is the solution of a heat equation, it is a $\mathcal{C}^\infty$ function far from $\theta = \pi$ for all time. As a consequence, symmetry also implies $\partial_\theta s(t, 0) = 0$ for all t. Hence, we reduce our study to the half-circle $\theta \in [0, \pi]$ and consider the following dynamics:

$$\begin{cases} \partial_t r = s, \\ \partial_t s = \Delta_r s, \\ s(t, \theta = \pi) = u(t), \\ \partial_\theta s(t, \theta = 0) = 0. \end{cases} \tag{2.27}$$

We now study the Riemannian structure on the cell induced by a shape r. As already stated, s is $\mathcal{C}^\infty$ except in 0, since its value there depends on $u(t)$. Assuming that the choice of u implies that s is $\mathcal{C}^\infty$ at 0 too, we have that r is a $\mathcal{C}^\infty$ function too. Consider the coordinate θ on the circle, and observe that a displacement ∂_θ on the coordinate induces a displacement in the r variable that can be estimated by $\sqrt{r^2 + r_\theta^2}\partial_\theta$, where $r_\theta = \partial_\theta r$ is the derivative of r with respect to θ. The estimate is due to a simple geometric first-order estimate of the length of the curve $r(\theta)$. As a consequence,

one can define the following metric on S^1:

$$g_\theta \text{ is bilinear and satisfies: } g_\theta(\partial_\theta, \partial_\theta) = r^2(\theta) + r_\theta^2(\theta). \tag{2.28}$$

This uniquely defines the metric on S^1. It is also clear that the inverse of the metric satisfies $g^\theta(d\theta, d\theta) = \frac{1}{r^2(\theta)+r_\theta^2(\theta)}$. Such an operator is never zero since the radius is supposed to be positive for all θ. Then, a direct computation gives the explicit expression of the Laplace-Beltrami operator Δ_r. We have:

$$\begin{aligned}
\Delta_r s &= \frac{1}{\sqrt{|g_\theta|}}\partial_\theta\left(\sqrt{|g_\theta|}g^\theta\partial_\theta s\right) \\
&= \frac{1}{\sqrt{r^2+r_\theta^2}}\partial_\theta\left(\frac{1}{\sqrt{r^2+r_\theta^2}}\partial_\theta s\right) \\
&= \frac{1}{r^2+r_\theta^2}\partial_\theta^2 s - \frac{rr_\theta + r_\theta\partial_\theta^2 r}{(r^2+r_\theta^2)^2}\partial_\theta s.
\end{aligned} \tag{2.29}$$

Hence the system we want to study is the following:

$$\begin{cases}
\partial_t r = s, \\
\partial_t s = \frac{1}{r^2+r_\theta^2}\partial_\theta^2 s - \frac{rr_\theta + r_\theta\partial_\theta^2 r}{(r^2+r_\theta^2)^2}\partial_\theta s, \\
s(t, \theta = \pi) = u(t), \\
\partial_\theta s(t, \theta = 0) = 0.
\end{cases} \tag{2.30}$$

We want to prove controllability for system (2.30) in a specific case, that is to find a control u that drives a (symmetric) cell shape to another (symmetric) cell shape in a given time interval $[0, T]$, together with having a signal s that is zero at the initial and final times. In mathematical terms, we consider initial and final configurations r_0, r_1 and a time $T > 0$. We want to find a control $u : [0; T] \to \mathbb{R}$ such that the unique solution of (2.30) with $r(t = 0) = r_0$ and $s(t = 0) = 0$ satisfies $r(t = T) = r_1$ and $s(t = T) = 0$. This goal is called *exact controllability*. It is known that this goal is impossible to be achieved in general, since we already know that some configurations (for instance non-smooth final configurations) cannot be reached with a heat equation.

Hence, we instead aim to prove *approximate controllability*, defined as follows: considering initial and final configurations r_0, r_1 and a time $T > 0$, for every $\epsilon > 0$, we want to find a control $u : [0; T] \to \mathbb{R}$ such that the unique solution of (2.30) with $r(t = 0) = r0$ and $s(t = 0) = 0$ satisfies $\|r(t = T) - r_1\|_{L^2} < \epsilon$ and $\|s(t = T)\|_{L^2} < \epsilon$.

It was shown in [69] that the 1-D generalized heat equation

$$\begin{cases} \partial_t \phi = f(\theta)\partial_\theta^2 \phi + g(\theta)\partial_\theta \phi + h(\theta)\phi, \\ \phi(t, \theta = \pi) = u(t), \\ \partial_\theta \phi(t, \theta = 0) = 0 \end{cases} \tag{2.31}$$

is approximately controllable where $f > 0$, g and h are analytic functions. Moreover, [69] proves a stronger condition: (approximate) motion planning or (approximate) tracking, defined as follows. Given a reference trajectory, we want to find a control such that the solution of the system (2.31) stays close to the reference trajectory for each time. In mathematical terms, one has the following result.

Theorem 2.3.1. *Consider a time horizon $[0, T]$ and a smooth trajectory $\bar{f} : [0, T] \to L^2(0, \pi)$. For every $\epsilon > 0$, there exists $u : [0, T] \to \mathbb{R}$ such that the solution of* (2.31) *with initial data $\bar{f}(0)$ satisfies $\|f(t) - \bar{f}(t)\|_{L^2} < \epsilon$ for all time $t \in [0, T]$.*

We use this result to prove approximate controllability of (2.30). Moreover, we will show a stronger condition, that is approximate tracking of the r variable, together with the condition $\|s(t = 0)\|_{L^2} < \epsilon$ and $\|s(t = T)\|_{L^2} < \epsilon$. Since we need analytic coefficients for the second equation of (2.30), we need a reference trajectory that is analytic for all t, i.e. $r : [0; T] \to C^w(0, \pi)$, together with smoothness with respect to t. We can prove the following main theorem.

Theorem 2.3.2. *Let $\bar{r} : [0, T] \to C^w(0, \pi)$ be a reference trajectory. Then for all $\epsilon > 0$, there exists a control $u : [0, T] \to \mathbb{R}$ such that the unique solution of* (2.30) *with $r(t = 0) = \bar{r}(t = 0)$ and $s(t = 0) = 0$ satisfies $\|r(t) - \bar{r}(t)\|_{L^2} < \epsilon$ for all $t \in [0, T]$.*

Proof. (Sketch). First observe that an approximate tracking of $\partial_t \bar{r}$ by s implies approximate tracking of $\bar{r}$ by r. Indeed, one has

$$\begin{aligned} \|\bar{r}(t) - r(t)\|_{L^2} \leq & \|\bar{r}(0) - r(0)\|_{L^2} \\ & + \int_0^t \|\partial_t \bar{r}(\tau) - s(\tau)\|_{L^2} d\tau \\ \leq & 0 + t\epsilon \leq T\epsilon. \end{aligned}$$

For simplicity, we define $\bar{s} = \partial_t \bar{r}$. We now prove approximate tracking of $\bar{s}$ by s. There are two main differences between our system (2.30) and the generalized heat equation (2.31) used in Theorem 2.3.2. Firstly, we require the bounds $\|s(t = 0)\|_{L^2} < \epsilon$ and $\|s(t = T)\|_{L^2} < \epsilon$, while $\bar{s}$ does not satisfy

this condition. Secondly, coefficients f, g, h in (2.31) do not depend on the solution ϕ. Instead, in our case the coefficients in (2.30) depend on r, which itself depends on s. The first difficulty can be overcome by tracking the following trajectory. Given $\eta > 0$, consider

$$\tilde{s}(\tau) := \begin{cases} \frac{\tau}{\eta}\bar{s}(\tau) & \tau \in [0, \eta[, \\ \bar{s}(\tau) & \tau \in [\eta, T-\eta[, \\ \frac{T-\tau}{\eta}\bar{s}(\tau) & \tau \in [T-\eta, T[\end{cases}$$

The definition of $\tilde{s}$ satisfies the conditions $\|\tilde{s}(t=0)\|_{L^2} = \|\tilde{s}(t=T)\| = 0$, hence the tracking of $\tilde{s}$ by s will give the conditions $\|s(t=0)_{L^2} = 0 < \epsilon$ and $\|s(t=T)\|_{L^2} < \epsilon$. The corresponding solution $\tilde{r}$ satisfies

$$\|\tilde{r}(t) - \bar{r}(t)\|_{L^2} \leq \int_0^T \|\tilde{s}(\tau) - \bar{s}(\tau)\| d\tau \leq \max_{t\in[0,T]} \|\bar{s}(t)\|_{L^2}.$$

Then, a choice of η sufficiently small gives the tracking of $\bar{r}$ by $\tilde{r}$, hence of $\bar{r}$ by r.

The second difficulty can be overcome by a sample-and-hold method for the second equation of (2.30). For a given natural number $n \in \mathbb{N}$, define $\{t_1; ...t_n\}$ as $t_k := \frac{k}{n}T$ and consider the following equation for $t \in [t_k; t_{k+1}[$:

$$\partial_t s = \frac{1}{r^2(t_k) + r_\theta^2(t_k)}\partial_\theta^2 s - \frac{r(t_k)r_\theta(t_k) + r_\theta(t_k)\partial_\theta^2 r(t_k)}{(r^2(t_k) + r_\theta^2(t_k))^2}\partial_\theta s \tag{2.32}$$

For this equation one can directly use the results in [69] for each time interval, since for each interval the coefficients in (2.32) are analytic functions, not depending on s. Then, given the reference trajectory $\tilde{s}$, one can iteratively solve the tracking problem as follows:

- Find the control $u^n \in [0, t_1[$ such that the solution s^n of (2.32) satisfies $\|\tilde{s}(\tau) - s^n(\tau)\|_{L^2} < \epsilon$ for all $t \in [0, t_1[$, by using Theorem 2.3.2.
- Compute $r^n(t)$ as the solution of $\partial_t r^n = s^n$ with $r(t=0) = r_0$ for $t \in [0, t_1]$. Observe that it is analytic.
- Plug $r^n(t_1)$ in (2.32) and find the control $u^n \in [t_1, t_2[$ such that the solution s^n of (2.32) satisfies $\|\tilde{s}(\tau) - s^n(\tau)\|_{L^2} < \epsilon$ for all $t \in [t_1, t_2[$, by using Theorem 2.3.2.
- Continue until $t^n = T$.

This iterative method provides a control $u^n : [0, T] \to \mathbb{R}$ such that the solution s^n of 2.32 satisfies $\|\tilde{s}(\tau) - s^n(\tau)\|_{L^2} < \epsilon$ for all time $t \in [0, T]$. We plug such control $u_n := u^n$ into 2.30 and we find a pair (r_n, s_n) that is somehow close to (r^n, s^n) for big n. We need a detailed estimate of such distance

(that is exponential), but there is no problem due to analyticity. Since s^n tracks $\tilde{s}$, then s_n tracks $\tilde{s}$ too, hence r_n tracks $\bar{r}$. $\square$

2.3.2 Equilibria

We look for equilibria of the form: $u(t) = u_e$ and $s(t,\theta) = s_e(\theta)$, that solves the system:

$$\begin{cases} \partial_t r_e = s_e, \\ \partial_\theta^2 s_e = \frac{r_e \partial_\theta r_e + \partial_\theta r_e \partial_\theta^2 r_e}{(r_e)^2 + (\partial_\theta r_e)^2} \partial_\theta s_e, \\ s_e(\theta = \pi) = u_e, \\ \partial_\theta s_e(\theta = 0) = 0. \end{cases} \tag{2.33}$$

From (2.30), we deduce that for all θ, $r_e(t,\theta)$ is a linear function of $s_e(\theta)$:

$$r_e(t,\theta) = s_e(\theta)t + r_0(\theta), \tag{2.34}$$

where $r_0(\theta) := r_e(0,\theta)$. One obvious possible equilibrium is obtained when there is no control, i.e. for a zero signal (since s_e then solves a Laplace equation with the boundary condition $s_e(\theta = \pi) = 0$). One gets:

$$\begin{cases} u_e = 0, \\ s_e(\theta) = 0 \quad \text{for all } \theta \in [0,\pi], \\ r_e(t,\theta) = r_0(\theta) \quad \text{for all } t \in [0,T], \text{ for all } \theta \in [0,\pi]. \end{cases}$$

Hence, if s_e and u_e are at an equilibrium such that $u_e = 0$, there is no signal and the radius is constant in time.

On the other hand, if $u_e > 0$, then s_e solves a Laplace-type equation with a non-zero Dirichlet boundary condition at $\theta = \pi$, so $s_e(\theta) > 0$ for all $\theta \in [0,\pi]$. Hence $r_e(t,\theta)$ grows linearly with time and does not reach an equilibrium. We instead look for an equilibrium in the shape of the membrane, by defining $\rho_e(t,\theta) = \frac{r_e(t,\theta)}{r_e(t,\theta=\pi)}$ (notice that this is possible since $r_e \neq 0$). Then ρ_e is constant in time if $\partial_t \rho_e = 0$, which gives:

$$\partial_t r_e(t,\theta) r_e(t,\pi) - \partial_t r_e(t,\pi) r_e(t,\theta) = 0.$$

Since $\partial_t r_e(t,\theta) = s_e(\theta)$, we get:

$$\partial_t \rho_e(t,\theta) = 0 \iff \frac{r_e(t,\theta)}{r_e(t,\pi)} = \frac{s_e(\theta)}{s_e(\pi)}.$$

This means that at each time t, the membrane r_e is a dilation of the signal s_e. In particular, at $t = 0$, $r_0(\theta) = \frac{r_0(\pi)}{s_e(\pi)} s_e(\theta)$ for all θ. Hence from (2.34) we get: $r_e(t,\theta) = s_e(\theta)(t + \frac{r_0(\pi)}{s_e(\pi)})$. Since $s_e(\theta)$ and $r_e(t,\theta)$ are proportional, the second equation of (2.33) becomes:

$$\partial_\theta^2 s_e = \frac{s_e \partial_\theta s_e + \partial_\theta s_e \partial_\theta^2 s_e}{(s_e)^2 + (\partial_\theta s_e)^2} \partial_\theta s_e,$$

which, after simplification, gives:

$$s_e \partial_\theta^2 s_e = (\partial_\theta s_e)^2.$$

One solution to this nonlinear differential equation is the constant signal $s_e(\theta) = u_e$, where s_e satisfies both the Neumann and Dirichlet boundary conditions prescribed in (2.33).

We relax our conditions and look for a solution s_e that satisfies $s_e(\pi) = u_e$ but not $\partial_\theta s_e(0) = 0$. In particular, if we suppose that $\partial s_e(\theta) \neq 0$ for all $\theta \in [0, T]$, we can write:

$$\frac{\partial_\theta^2 s_e}{\partial_\theta s_e} = \frac{\partial_\theta s_e}{s_e}.$$

Then $\partial_\theta(\ln(\partial_\theta s_e)) = \partial_\theta(\ln(s_e))$, so we get: $s_e(\theta) = u_e e^{\lambda(\theta-\pi)}$, where λ is a constant. Notice that then we can bring $\partial_\theta s_e(0) = u_e \lambda e^{-\lambda\pi}$ arbitrarily close to zero by choosing λ, so we partially recover the original Neumann boundary condition.

2.3.3 Simulations

We simulate diffusion of the signal by discretizing the second equation of system (2.30) using Finite Differences, supplemented by a Neumann boundary condition at angle $\theta = 0$ ($\partial_\theta s(t,0) = 0$) and a Dirichlet boundary condition at angle $\theta = \pi$ ($s(t,\pi) = u(t)$). Then the radius of the manifold at each time-step is obtained by simple integration of the signal.

Comparison of diffusion on a static vs growing manifold

We run simulations for a constant control $u_1 \equiv 1$, an initial signal $s_0(\theta) = 0$ and an initial radius $r_0(\theta) = 1$ for all $\theta \in [0, \pi]$. We notice that s reaches an equilibrium after time $t = 2$. After that point, the radius grows in a linear way, i.e. $\rho(t) = \text{const}$. See Figure 2.5.

We then turn our attention to the comparison with the case in which we neglect the growth of the manifold (this would correspond to an egg chamber of constant size). In this case, taking as initial condition a circle, the radius r is constant both w.r.t. time and the θ variable, thus $r \equiv 1$ and $r_\theta \equiv 0$. Plugging this information into equation (2.30), the Laplace-Beltrami operator reduces to standard diffusion and we get the following system:

$$\begin{cases} \partial_t r = s, \\ \partial_t s = \partial_\theta^2 s \\ s(t, \theta = \pi) = u(t), \\ \partial_\theta s(t, \theta = 0) = 0. \end{cases} \tag{2.35}$$

The simulations for a constant control $u \equiv 1$ are very different from those obtained by using the system (2.30): Figure 2.6 shows the evolution of the signal and the radius with constant control for system. The signal s reaches an equilibrium $s(t, \theta) = 1$, which means that the growth of the radius tends to be uniform with respect to the angle θ. Therefore, as expected, neglecting the growth of the manifold generates uniform growth and, in the biological system, would give rise to spherical egg chambers opposed to the spheroidal ones observed in nature.

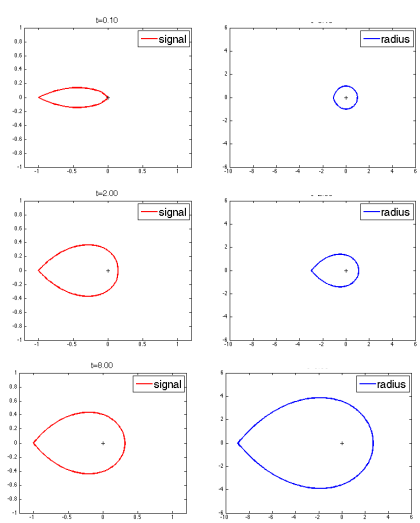


Figure 2.5: Signal s (left) and radius r (right) for a constant control $u \equiv 1$ at times $t = 0.1$, $t = 2$ and $t = 8$. The source correponds to the angle $\theta = \pi$, so in the signal picture it is located on the left end of the equator line corresponding to coordinates $(-1, 0)$.

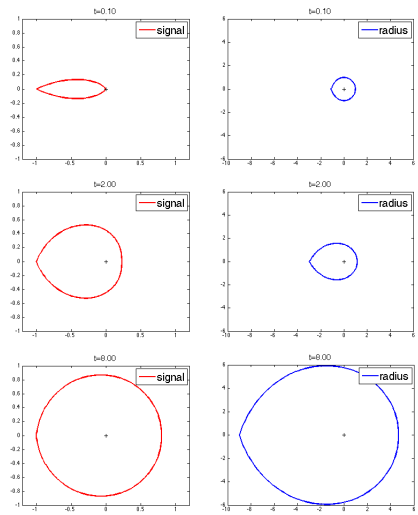


Figure 2.6: Signal s (left) and radius r (right) for a constant control $u \equiv 1$ at times $t = 1$, $t = 2$, $t = 5$ and $t = 8$.

Single source

A source placed on the first axis (at angle $\theta = \pi$) allows us to control the diameter of the manifold along the same axis. In Figure 2.5, the manifold is stretched along the first axis direction at final time, with an emphasis on the left side, i.e. $r(T, \pi) > r(T, 0)$. Using the source to impose negative values of the signal (which has a mathematical meaning but not a biological one), we can control the final shape of the manifold to achieve $r(T, \pi) < r(T, 0)$. In order to do that we set the control as:

$$u(t) = \begin{cases} 0.5 \cdot sin(\omega t) & t \in [0, 5] \\ 0 & t \in]5, 10] \end{cases},$$

where $\omega = \frac{2\pi}{5}$ so that we obtain a complete sinusoidal oscillation up to time 5 then the signal is vanishing (which coincides with control u_2 depicted in Figure 2.7). The final result is a apple shape manifold with pitch located at the signal source point, see Figure 2.8. To better visualize the relationship between the signal and the shape we visualized the signal on the manifold itself, so for positive values the signal will be outside the manifold and inside for negative ones.

Using a single source it is also possible to induce an homogeneous growth along all directions, but with time-dependent signals. We first give an impulse and then turn off the signal. Define the control by:

$$u(t) = \begin{cases} 0.2 \cdot sin(\omega t) & t \in [0, 2.5] \\ 0 & t \in]5, 10] \end{cases}, \tag{2.36}$$

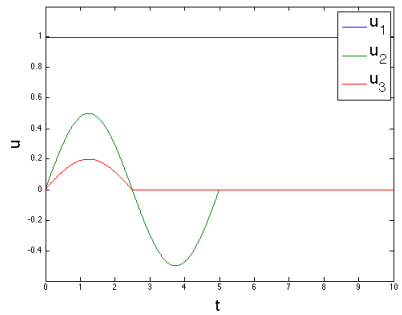


Figure 2.7: Control functions u_1, u_2 and u_3

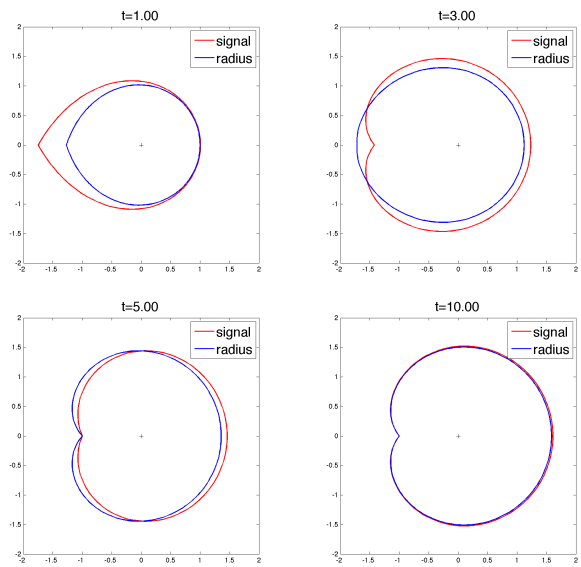


Figure 2.8: Radius r (in blue) and signal s (plotted as $r + s$ in red) for a control $u = u_2$ at times $t = 1, 3, 5$, and 10.

where $\omega = \frac{2\pi}{5}$, so that the half sinusoidal oscillation gives an always positive signal (this correspond also to the control u_3 depicted in Figure 2.7). The final shape is close to that of a circle, but with a larger radius than that at initial time (see Figure 2.9).

Double source

As observed above, a single static source allows us to control the radii $r(T, 0)$ and $r(T, \pi)$, i.e. the horizontal growth. In order to achieve a larger growth along the vertical axis, we consider a system with double source: one locate at angle $\theta = 0$ and the second (as before) at angle $\theta = \pi$. We obtain

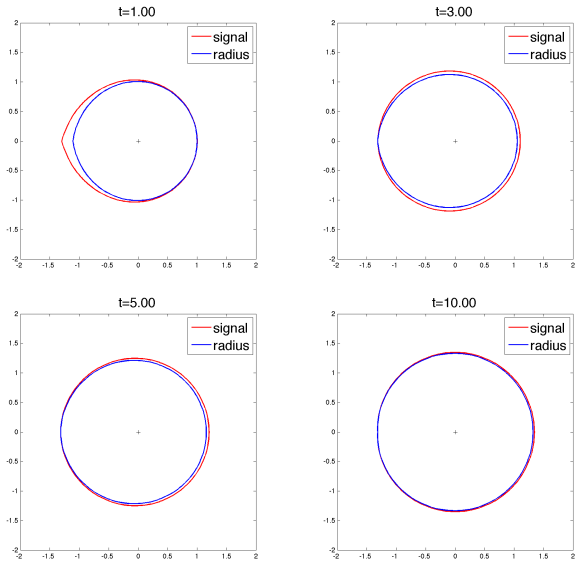


Figure 2.9: Radius r (in blue) and signal s (plotted as $r+s$ in red) for a control $u = u_3$ at times $t = 1, 3, 5$ and 10.

the system:

$$\begin{cases} \partial_t r = s_L + s_R, \\ \partial_t s_L = \frac{1}{r^2+r_\theta^2}\partial_\theta^2 s_L - \frac{rr_\theta + r_\theta \partial_\theta^2 r}{(r^2+r_\theta^2)^2}\partial_\theta s_L \\ \partial_t s_R = \frac{1}{r^2+r_\theta^2}\partial_\theta^2 s_R - \frac{rr_\theta + r_\theta \partial_\theta^2 r}{(r^2+r_\theta^2)^2}\partial_\theta s_R \\ s_L(t, \theta = \pi) = u_L(t), \quad s_R(t, \theta = 0) = u_R(t), \\ \partial_\theta s_L(t, \theta = 0) = 0, \quad \partial_\theta s_R(t, \theta = \pi) = 0. \end{cases} \tag{2.37}$$

If we use the control given by formula (2.36) for both sources, we obtain a final manifold stretched more in the vertical direction, i.e. $r(T, \pi/2) > r(T, 0) = r(T, \pi)$, see Figure 2.10.

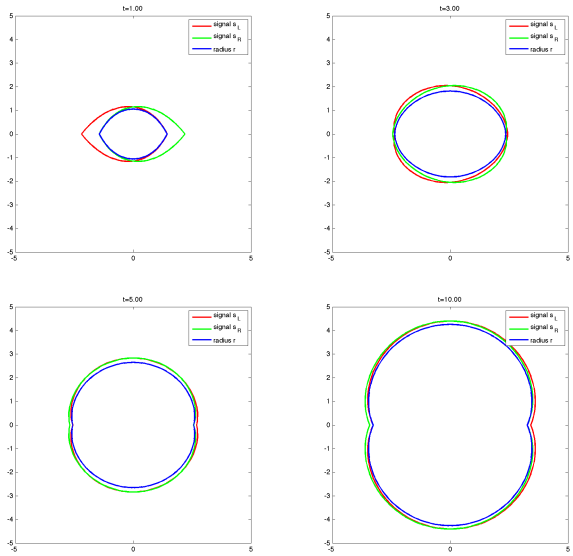


Figure 2.10: Radius r (in blue) and signals s_L (plotted as $r + s_L$ in red) and s_R (plotted as $r + s_R$ in green) for controls $u_L = u_R = u_3$ at times $t = 1$, $t = 3$, $t = 5$ and $t = 10$.

Part II

Social dynamics models

Chapter 3

Achieving consensus: Optimal control of a collective migration model

Introduction

A fascinating feature of large groups is their *self-organization* ability, i.e. the emergence from local interaction rules of certain global patterns. For instance, animal groups such as schools of fish, flocks of birds or herds of mammals exhibit strong coordination in their movements [7, 8, 18, 24, 25, 87, 88, 89, 103, 124, 128]. This collective behavior in animal groups also inspired applications to robotics [9], in which the aim is to coordinate autonomous vehicles[23, 60, 74, 121] and flight formations [91, 111]. Other interests concern models in microbiology [57, 58, 61, 90, 93], pedestrian and crowd motions[29, 30] and financial markets [4, 37, 70]. Such systems are usually referred to as social dynamics. Examples of self-organization include clustering of the agents, alignment of velocities, or other kinds of equilibria [16, 50, 82, 87, 88, 89, 124]. This raises the question of understanding the mechanisms behind the global pattern formation.

A well-known model was proposed by F. Cucker and S. Smale [31] to describe the phenomenon of *consensus* in terms of alignment of velocities in a group on the move. The Cucker-Smale model in formula is written as:

$$\begin{cases} \dot{x_i} = v_i \\ \dot{v_i} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} \dfrac{v_j - v_i}{(1 + \|x_j - x_i\|^2)^\beta} \end{cases} \qquad \text{for } i \in \{1, ..., N\}, \tag{3.1}$$

where $\beta > 0$, and $x_i \in \mathbb{R}^d$ and $v_i \in \mathbb{R}^d$ are respectively the *state* and *velocity*. This model was originally designed to describe the formation and evolution of language, and the variables v_i can more generally represent opinions, preferences or invested capital. The system converges to consensus

if $\beta \leq \frac{1}{2}$, which corresponds to a strong interaction even between distant agents [18, 18]. On the other hand, if $\beta > \frac{1}{2}$, i.e. if the interaction is too weak, convergence to consensus only happens under certain conditions. More generally, the term $(1 + \|x_j - x_i\|^2)^{-\beta}$ can be replaced by $a(\|x_j - x_i\|)$. Intuitively, it is natural to define a as a non-increasing function, since proximity often encourages interaction. On the other hand, it was proven that interactions modeled by non-decreasing functions a, called heterophilious, in fact enhance consensus [82]. When the system does not converge to a desired state, a natural question is to study the possibility of steering it via controls functions u_i, in which case the second equation of (3.1) becomes: $\dot{v_i} = \frac{1}{N}\sum_{j=1}^{N} a(\|x_j - x_i\|)(v_j - v_i) + u_i$ [17, 18, 39].

In the *collective migration* problem [75], not only do agents interact with one another to travel as a group, but they also gather clues from the environment guiding them towards a global *target velocity*. In the case of migrating birds, for instance, this velocity can be sensed through a magnetic field, the direction of the sun, or environmental features. However, sensing the migration velocity is costly, both in used time and energy. A trade-off thus occurs between gathering this information, which ensures more precision, and following the group, which is less costly and saves time and energy for other tasks such as surveying for predators [32, 48]. This problem also applies to the field of robotics, in which gathering information from the environment is done at the expense of communicating with other robots (or planes, drones, etc.) or performing other tasks, and to the field of economics when one aims to influence decisions of a group based on limited information. This trade-off naturally separates the group into *leaders*, who gather information, and *followers*, who only interact with the other agents [48]).

We study a *Collective Migration Model*, where the agents' dynamics is determined by two forces: the attraction towards a target velocity V (which we assume can be sensed) and the consensus dynamics as in the Cucker-Smale model. More precisely, each agent's evolution is governed by a parameter $\alpha_i \in [0, 1]$ which provides the balance between the two forces. The system can be written as:

$$\begin{cases} \dot{x_i} = v_i \\ \dot{v_i} = \alpha_i(V - v_i) + (1 - \alpha_i)\dfrac{1}{N}\displaystyle\sum_{j=1}^{N} a(\|x_j - x_i\|)(v_j - v_i) \end{cases} \quad \text{for } i \in \{1, ..., N\}, \tag{3.2}$$

where $x_i \in \mathbb{R}^d$ and $v_i \in \mathbb{R}^d$ are the state and velocity, $V \in \mathbb{R}^d$ is the target velocity, and $\alpha_i \in [0, 1]$ is the control, with the constraint $\sum_i \alpha_i \leq M$, $M > 0$. Here, we choose to set $a \equiv 1$, so that the strength of interaction does not depend on the agents' positions. This is a reasonable hypothesis for instance if we consider groups of planes or drones that can communicate just as easily from great

distances.

While the Cucker-Smale model leads to alignment of all velocities to the average one (when there is consensus), the migration model tends to align all velocities to the preassigned *target velocity.* Our work focuses on finding optimal control strategies in order to achieve consensus to the target velocity, and in particular on selecting optimal *controlled leaders* among the agents when the control strength M is small with respect to the size of the group. In order to do that, we define the cost function $\tilde{\mathbb{V}} = \frac{1}{N}\sum_i \|v_i - V\|^2$, measuring the distance from consensus at the target velocity. We first show that, given any $M > 0$, the strategy to decrease $\tilde{\mathbb{V}}$ instantaneously, with the constraint $\sum_i \alpha_i \leq M$, consists of distributing the control among the agents with the largest positive projections of velocities along $\bar{v} - V$ (where $\bar{v}$ is the mean velocity). In particular, if $\langle v_i, \bar{v} - V\rangle < 0$, the agent i is not controlled ($\alpha_i = 0$).

We then study the optimal control strategy to minimize $\tilde{\mathbb{V}}$ at a fixed final time and first focus on the case of two agents, with control bounded by $M \in [0, 2]$. The optimal control strategies depend on M but, in all cases, we act with larger control on the agent with the largest projected velocity. Furthermore, if the final time is too short to bring the agents together, then there are initial conditions for which at first the system must evolve with no control ($\alpha \equiv 0$). We call this phenomenon “Inactivation", in line with the “Inactivation Principle" proven in [10] in the context of arm movements. In this collaborative work with biologists, the authors prove that during fast arm movements, it is optimal to simultaneously inactivate both agonistic and antagonistic muscles for a short moment nearing the peak velocity. We next generalize our results to any number of agents, but with the constraint $M \leq 1$. Then the optimal control strategy acts with full strength on a sub-group of agents to bring them together. Also in this case we observe “Inactivation", which occurs when the initial average velocity $\bar{v}$ is very close to the target velocity V. Indeed, driving the system to V requires both achieving consensus and moving the average velocity towards V. If the average velocity is already close to V, then we are left with inducing consensus which happens naturally without control. However, simulations show that Inactivation is rare and its performance gain is very minor compared to a full-control strategy.

Then we move on to examine integral costs $\int_0^T \tilde{\mathbb{V}}(t)dt$ and show that the optimal control strategy never exhibits Inactivation. More precisely, we must use full control at all time splitting it evenly among the agents with the biggest projected velocity. Such a strategy is more restrictive than that with final cost, since the controls are completely determined by initial conditions, while previously we could use any strategy bringing agents together at final time.

This chapter is organized as follows. In Section 3.1, we define the cost functional and make

general observations. In Section 3.2, we determine the strategy to decrease it instantaneously in time. Then, in Section 3.3, we introduce the optimal control problem to minimize the cost function at a given final time. We solve it for the particular case of two agents (Section 3.4) before generalizing to any number of agents with a control bounded by 1 (Section 3.5). Lastly we find optimal control strategies to minimize the integral cost (Section 3.6).

3.1 Cost function and general observations

With no loss of generality, we set the target velocity V to zero. Having simplified the interaction function a, system (3.2) reduces to:

$$\begin{cases} \dot{x_i} = v_i \\ \dot{v_i} = -\alpha_i v_i + (1-\alpha_i)\dfrac{1}{N}\sum\limits_{j=1}^{N}(v_j - v_i) \end{cases} \qquad i \in \{1, ..., N\}. \tag{3.3}$$

We set a final time $T > 0$. Then given $M > 0$, we define the set of controls $\mathcal{U}_M$ as:

$$\mathcal{U}_M = \left\{ \alpha : [0,T] \to [0,1]^N \middle| \ \alpha \text{ measurable, s.t. for all } t, \ \sum_{i=1}^{N} \alpha_i(t) \leq M \right\}. \tag{3.4}$$

3.1.1 Projection of the Dynamics

Note that the dynamics (3.3) can be written in the more compact way:

$$\begin{cases} \dot{x_i} = v_i \\ \dot{v_i} = -v_i + (1-\alpha_i)\,\bar{v}, \end{cases} \tag{3.5}$$

where $\bar{v}$ represents the mean velocity $\bar{v} = \frac{1}{N}\sum_i v_i$. The evolution of $\bar{v}$ is given by $\dot{\bar{v}} = -\frac{1}{N}(\sum_i \alpha_i)\bar{v}$, so the direction of $\bar{v}$ is an invariant of the dynamics. We begin by assuming that the initial average velocity is different from the target one:

Hypothesis 1. $\bar{v}(0) \neq 0$.

This first assumption is only made in order to render the problem interesting. Indeed, if $\bar{v}(0) = 0$, i.e. if the mean velocity is already at the target velocity V, then according to the evolution $\dot{\bar{v}} = -(\sum_i \alpha_i)\bar{v}$, it would hold $\bar{v}(t) = 0$ for all $t \geq 0$. Then looking at Equation (3.5), we notice that the system is not controllable and that each velocity decreases exponentially to zero. We can then

define the invariant unit vector $e = \frac{\bar{v}}{\|\bar{v}\|}$.

Let $w_i = v_i - \langle v_i, e\rangle\, e$ be the projection of v_i over $(\bar{v}^{\perp})$. Then

$$\dot{w}_i = -v_i + (1-\alpha_i)\,\bar{v} - \langle -v_i + (1-\alpha_i)\,\bar{v}, e\rangle\, e = -w_i. \tag{3.6}$$

Therefore the projection of v_i over $(\bar{v}^{\perp})$ decreases exponentially, independently of the controls α_i. Let us now define $\xi_i = \langle v_i, e\rangle$. Its evolution is given by: $\dot{\xi}_i = -\langle v_i, e\rangle + (1-\alpha_i)\langle \bar{v}, e\rangle = -\xi_i + (1-\alpha_i)\|\bar{v}\| = -\xi_i + (1-\alpha_i)\bar{\xi}$. In the following, we will only study the equations governing the evolution of the projected variables ξ_i:

$$\text{For all } i \in \{1, ..., N\}, \quad \dot{\xi}_i = -\xi_i + (1-\alpha_i)\bar{\xi}, \tag{3.7}$$

where $\bar{\xi} = \frac{1}{N}\sum_j \xi_j$. This is a significant result: instead of studying a system evolving in $\mathbb{R}^{Nd}$, we consider a system in $\mathbb{R}^N$, thus greatly reducing the complexity of theoretical and numerical analyses. Hereafter we shall make the following hypothesis:

Hypothesis 2. *$\xi_i(0) \geq \xi_{i+1}(0)$ for every $i \in \{1, ..., N-1\}$.*

This assumption allows us to order the initial projected velocities without loss of generality.

Proposition 3.1.1.

Having made Hyp. 1 and Hyp. 2, it holds $\bar{v}(t) \neq 0$ and $\bar{\xi}(t) > 0$ for all $t \in [0, T]$.

Furthermore, let $\tau \in [0, T]$. If $\xi_i(\tau) \geq 0$, then $\xi_i(t) \geq 0$ for all $t \in [\tau, T]$. If $\xi_i(\tau) > 0$, then $\xi_i(t) > 0$ for all $t \in [\tau, T]$.

Proof. The proposition is mainly a consequence of Gronwall's inequality: It holds

$$\bar{\xi} = \frac{1}{N}\sum_j \langle v_j, \frac{\bar{v}}{\|\bar{v}\|}\rangle = \langle \bar{v}, \frac{\bar{v}}{\|\bar{v}\|}\rangle = \|\bar{v}\| \tag{3.8}$$

and

$$\dot{\bar{\xi}} = -\frac{1}{N}\left(\sum_{i=1} \alpha_i\right)\bar{\xi} \geq -\frac{M}{N}\bar{\xi}.$$

Hence, if $\bar{v}(0) \neq 0$ and therefore $\bar{\xi}(0) > 0$, then $\bar{\xi}(t) \geq e^{-Mt/N}\bar{\xi}(0) > 0$ and thus $\bar{v}(t) \neq 0$ for all $t \in [0, T]$. Now notice that from (3.7) we can compute for all $t \in [\tau, T]$: $\xi_i(t) = e^{-(t-\tau)}(\xi_i(\tau) + \int_\tau^t (1-\alpha_i)(s)\bar{\xi}(s)e^{s-\tau}ds)$, so $\xi_i(t) \geq e^{-(t-\tau)}\xi_i(\tau)$, which proves the second part of the proposition. $\square$

3.1.2 Migration functional

We introduce the functional

$$\tilde{\mathbb{V}} = \frac{1}{N}\sum_{i=1}^{N} \|v_i - V\|^2, \tag{3.9}$$

which measures the distance from consensus at the desired velocity V. Since we set $V = 0$, $\tilde{\mathbb{V}}$ reduces to: $\tilde{\mathbb{V}} = \frac{1}{N}\sum_i \|v_i\|^2$. In the new projected coordinates ξ, the migration functional can be written as: $\tilde{\mathbb{V}} = \frac{1}{N}\sum_i(\|w_i\|^2 + \xi_i^2)$, where only the second term ξ_i^2 can be controlled. Hence, here onward we will only consider the controllable part of $\tilde{\mathbb{V}}$, which we denote $\mathbb{V}$:

$$\mathbb{V} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^2. \tag{3.10}$$

Notice that $\mathbb{V}$ can be written as a sum of two terms:

$$\mathbb{V} = \bar{\xi}^2 + \frac{1}{N}\sum_{i=1}^{N}(\xi_i - \bar{\xi})^2, \tag{3.11}$$

which should be minimized simultaneously (where we remind that $\bar{\xi} = \frac{1}{N}\sum_i \xi_i$). Minimizing $\bar{\xi}^2$ (or $\bar{\xi}$, since according to Proposition 3.1.1, $\bar{\xi} > 0$) corresponds to steering the system as a whole to the desired velocity $V = 0$. On the other hand, minimizing $\frac{1}{N}\sum_i(\xi_i - \bar{\xi})^2$ corresponds to driving the system to consensus. However, the dynamics (3.7) of ξ_i show that if $\xi_i < 0$, decreasing $\bar{\xi}$ slows down the increase of ξ_i, resulting in a possible increase of $(\xi_i - \bar{\xi})^2$. Hence, minimizing $\mathbb{V}$ requires balancing the decrease of the two terms in (3.11).

3.1.3 Minimization problems

In the following sections, we will deal with the minimization of different quantities, in order to design a strategy for consensus at the migration velocity $V = 0$. Having fixed the final time T a priori, we address three problems:

(i) The minimization of $\frac{d\mathbb{V}}{dt}$, i.e. the maximization of the instantaneous decrease of $\mathbb{V}$ (see Section 3.2).

(ii) The minimization of the final cost $\mathbb{V}(T)$ (see Sections 3.3, 3.4 and 3.5).

(iii) The minimization of the integral cost $\int_0^T \mathbb{V}(t)dt$ (see Section 3.6).

In order to minimize (ii) $\mathbb{V}(T)$ and (iii) $\int_0^T \mathbb{V}(t)dt$, we will design an optimal control strategy using Pontryagin's maximum principle. The minimization of $\dot{\mathbb{V}}$, on the other hand, will not provide

an optimal control.

3.2 Instantaneous Decrease

In this section we look for a control strategy maximizing the instantaneous decrease of $\mathbb{V}$. Strategies designed in this way are not optimal (in general), but are easier to study and can give a first good insight on the problem. Indeed, we will later compare the instantaneous decrease strategy to the optimal control strategies developed in Sections 3.5 and 3.6.

The time derivative of the migration functional $\mathbb{V}$ is given by:

$$\dot{\mathbb{V}} = \frac{2}{N}\sum_{i=1}^{N}\xi_i\dot{\xi}_i = \frac{2}{N}\left(\sum_{i=1}^{N}-\xi_i^2 + \sum_{i=1}^{N}(1-\alpha_i)\bar{\xi}\xi_i\right) = -2\mathbb{V} + \frac{2}{N}\bar{\xi}\sum_{i=1}^{N}(1-\alpha_i)\xi_i. \tag{3.12}$$

Since $\bar{\xi} \geq 0$, minimizing $\dot{\mathbb{V}}$ amounts to the following problem:

$$\text{Find} \quad \min \sum_{i=1}^{N}(1-\alpha_i)\xi_i, \tag{3.13}$$

which can be done as follows (where $\lfloor M \rfloor$ and $\lceil M \rceil$ respectively denote the floor and the ceiling of M, and $|\cdot|$ denotes the cardinality of a set):

Proposition 3.2.1. *Suppose that $\xi_1(t) \geq ... \geq \xi_N(t)$ (or re-arrange the agents so that this is satisfied). Then the following strategy minimizes $\frac{d}{dt}\mathbb{V}$ at time t:*
Define $I^+(t) = \{i \in \{1,..,N\},\ \xi_i(t) > 0\}$.
If $|I^+(t)| \leq M$, then set $\alpha_i(t) = 1$ if $i \in I^+$ and $\alpha_i(t) = 0$ otherwise.
If $|I^+(t)| > M$ and $\xi_{\lceil M-1 \rceil} > \xi_{\lceil M \rceil} > \xi_{\lceil M+1 \rceil}$ then set $\alpha_i(t) = 1$ if $i \leq \lfloor M \rfloor$, $\alpha_{\lfloor M \rfloor+1}(t) = M - \lfloor M \rfloor$ and $\alpha_i(t) = 0$ otherwise.
If $|I^+(t)| > M$ and $\xi_{\lceil M-1 \rceil} = \xi_{\lceil M \rceil}$ or $\xi_{\lceil M \rceil} = \xi_{\lceil M+1 \rceil}$, let $I_{\lceil M \rceil} = \{i \in \{1,...,N\},\ \xi_i(t) = \xi_{\lceil M \rceil}(t)\}$ and $I^*_{\lceil M \rceil} = \{1,...,\lceil M \rceil\} \setminus I_{\lceil M \rceil}$. Then set $\alpha_i(t) = 1$ if $i \in I^*_{\lceil M \rceil}$, $\alpha_i(t) = \frac{M - |I^*_{\lceil M \rceil}|}{|I_{\lceil M \rceil}|}$ if $i \in I_{\lceil M \rceil}$ and $\alpha_i(t) = 0$ otherwise.

3.3 Optimal control for final cost

In this section, we focus on problem (ii) (see Section 3.1.3), i.e. minimizing the migration functional $\mathbb{V}$ at final time T using Pontryagin's maximum principle.

Let us compute the Hamiltonian H of the scalar system (3.7):

$$H = \sum_{i=1}^{N} \lambda_i \left(-\xi_i + (1-\alpha_i)\bar{\xi}\right) = -\bar{\xi}\sum_{i=1}^{N} \alpha_i\lambda_i + \sum_{i=1}^{N} \lambda_i \left(-\xi_i + \bar{\xi}\right). \tag{3.14}$$

By Pontryagin's maximum principle,[99] if $\alpha \in \mathcal{U}_M$, associated with the trajectory ξ, is optimal on $[0,T]$, then there exists $\lambda : [0,T] \to \mathbb{R}^N$ such that $\dot{\xi} = \frac{\partial H}{\partial \lambda}$ and $\dot{\lambda} = -\frac{\partial H}{\partial \xi}$. Furthermore the following minimization condition holds for almost all $t \in [0,T]$:

$$H(t,\xi(t),\lambda(t),\alpha(t)) = \min_{\beta \in \mathcal{U}_M} H(t,\xi(t),\lambda(t),\beta(t)). \tag{3.15}$$

Since $\bar{\xi} \geq 0$, minimizing H requires to set $\alpha_i = 1$ on the biggest positive λ_i. The differential equation for the covectors λ_i gives:

$$\dot{\lambda}_i = -\frac{\partial H}{\partial \xi_i} = \frac{1}{N}\sum_{j=1}^{N} \alpha_j\lambda_j - \bar{\lambda} + \lambda_i, \quad i \in \{1,...,N\}. \tag{3.16}$$

From this we can also compute the evolution of $\bar{\lambda} = \frac{1}{N}\sum_i \lambda_i$:

$$\dot{\bar{\lambda}} = \frac{1}{N}\sum_{j=1}^{N} \alpha_j\lambda_j. \tag{3.17}$$

Since the final condition for ξ is not fixed, the final condition for λ at time T gives:

$$\lambda(T) = \nabla\mathbb{V}(\xi(T)) = \left(\frac{2}{N}\xi_1(T),...,\frac{2}{N}\xi_N(T)\right). \tag{3.18}$$

Proposition 3.3.1. *If $\bar{t} > 0$, $i,j \in \{1,...,N\}$, and $\lambda_i(\bar{t}) = \lambda_j(\bar{t})$, then $\lambda_i(t) = \lambda_j(t)$ for all t. In this case, for a given control α, any control $\tilde{\alpha}$ satisfying $\tilde{\alpha_i} + \tilde{\alpha_j} = \alpha_i + \alpha_j$ and $\tilde{\alpha_k} = \alpha_k$ for every $k \neq i,j$ gives the same evolution of λ. If the control α satisfies the Pontryagin Maximum Principle, then the control $\tilde{\alpha}$ also does.*

Proof. Assume that at time $\bar{t}$, $\lambda_i(\bar{t}) = \lambda_j(\bar{t})$. Let us define $z_{ij} = \lambda_i - \lambda_j$. The evolution of z_{ij} is given by: $\dot{z}_{ij} = \dot{\lambda}_i - \dot{\lambda}_j = \lambda_i - \lambda_j = z_{ij}$. Hence, $z_{ij}(t) = z_{ij}(\bar{t})e^{t-\bar{t}}$, and if $z_{ij}(\bar{t}) = 0$, then for all t, $z_{ij}(t) = 0$, i.e. $\lambda_i(t) = \lambda_j(t)$. From this it follows that if α minimizes the Hamiltonian H, then any control $\tilde{\alpha}$ satisfying $\tilde{\alpha_i} + \tilde{\alpha_j} = \alpha_i + \alpha_j$ and $\tilde{\alpha_k} = \alpha_k$ also minimizes H, since one easily sees from (3.14) that $H^\alpha = H^{\tilde{\alpha}}$ (where we denote by H^α the Hamiltonian obtained with the control function α). $\square$

Still assuming that the projected velocities are initially ordered (Hypothesis 2), the following lemma will allow us to further assume that they are ordered at all time.

Lemma 3.3.1.

There exists an optimal strategy satisfying the following: For all $t \in [0,T]$,

$$\text{If } i < j, \text{ then } \xi_i(t) \geq \xi_j(t). \tag{3.19}$$

Proof. Consider an optimal control strategy $\alpha \in \mathcal{U}_M$.
Define $\tau = \sup\{t \in [0,T];\ \exists \beta \in \mathcal{U}_M \text{ s.t. } \mathbb{V}_\beta(T) = \mathbb{V}_\alpha(T) \text{ and } \xi^\beta \text{ satisfies (3.19) on } [0,t]\}$, where $\mathbb{V}_\beta$ and ξ^β denote respectively the migration functional and the dynamics driven by the control β. Let us prove by contradiction that $\tau = T$. Suppose that $\tau < T$. Then there exist $i, j \in \{1, ..., N\}$ with $i < j$ such that $\xi_i^\beta(\tau) = \xi_j^\beta(\tau)$ and $\xi_j^\beta(t) > \xi_i^\beta(t)$ on $]\tau, \tau+\delta]$ for some $\delta > 0$. Design a control strategy $\tilde{\beta}$ such that on $[\tau, T]$, $\tilde{\beta}_i = \beta_j$, $\tilde{\beta}_j = \beta_i$ and for every $k \in \{1, ..., N\} \setminus \{i, j\}$, $\tilde{\beta}_k = \beta_k$. Then for all $t \in [\tau, T]$, $\xi_i^{\tilde{\beta}}(t) = \xi_j^\beta(t)$, and $\xi_j^{\tilde{\beta}}(t) = \xi_i^\beta(t)$. So for all $t \in [\tau, \tau+\delta]$, $\xi_i^{\tilde{\beta}}(t) \geq \xi_j^{\tilde{\beta}}(t)$ and $\mathbb{V}^{\tilde{\beta}}(T) = \mathbb{V}^\beta(T)$. Proceeding likewise for every pair of indices (m, n) satisfying $m < n$ and $\xi_m^\beta(t) < \xi_n^\beta(t)$ on $]\tau, \tau+\delta]$ we are able to design a control strategy $\tilde{\beta}$ satisfying (3.19) on $[0, \tau+\delta]$ and $\mathbb{V}^{\tilde{\beta}}(T) = \mathbb{V}^\alpha(T)$, which contradicts the definition of τ. In conclusion, $\tau = T$, i.e. for all $t \in [0,T]$, for every $i, j \in \{1, ..., N\}$, if $i < j$ then $\xi_i(t) \geq \xi_j(t)$. □

Hence, from here onward we shall assume that the variables ξ_i are ordered at all time.

Hypothesis 3. *If $i < j$, then $\xi_i(t) \geq \xi_j(t)$ for all $t \in [0,T]$.*

From Hyp.3 and the transversality condition (3.18), we know that the covectors are ordered at final time, i.e. $\lambda_1(T) \geq ... \geq \lambda_N(T)$. From Prop. 3.3.1, we can generalize this for any time t:

$$\lambda_1(t) \geq ... \geq \lambda_N(t) \quad \text{for all } t \in [0,T]. \tag{3.20}$$

The Pontryagin Maximum Principle allows us to state the following:

Proposition 3.3.2. *The optimal strategy requires controlling the agents with the biggest positive covectors. Let $\alpha \in \mathcal{U}_M$ be an optimal strategy and λ_i, $i \in \{1, ..., N\}$ the corresponding covectors. Define:*

$$I_\lambda(t) := \Big\{ i \in \{1, ..., N\} \ \Big|\ \lambda_i(t) \geq 0 \Big\} \quad \text{and} \quad I_\lambda^+(t) := \Big\{ i \in \{1, ..., N\} \ \Big|\ \lambda_i(t) > 0 \Big\}. \tag{3.21}$$

If the set $I_\lambda(t)$ is empty, then there is no control on any agent: $\alpha_i(t) = 0$ for every i.
If the set $I_\lambda^+(t)$ is not empty, then there exists $i \in I_\lambda^+(t)$ such that $\alpha_i(t) > 0$. Furthermore, $\sum_j \alpha_j \geq \min(|I_\lambda^+(t)|, M)$.

Proof. According to Pontryagin's maximum principle (3.15), if the control α is optimal, then it minimizes the Hamiltonian H (3.14) for almost all $t \in [0, T]$. The only controllable part of H is $\tilde{H} = -\bar{\xi} \sum_i \alpha_i \lambda_i$. Minimizing H requires controlling the largest positive λ_i with the maximum strength allowed, while setting $\alpha_i = 0$ if $\lambda_i < 0$. If $\lambda_i = 0$, Pontryagin's maximum principle gives no information on α_i. □

This leads to a trichotomy of cases.

- The biggest positive λ_i's are always controlled with maximum control: $\sum_{i \in I_\lambda^+} \alpha_i = \min(|I_\lambda^+|, M)$.
- If for i, j, λ_i and λ_j coincide (at a certain time, which implies at all time) then α_i and α_j are under-determined. The PMP only requests that $\alpha_i + \alpha_j = c$ where c is given by the strength of the control to be used on the two agents.
- The negative λ_i's are never controlled: if $\lambda_i < 0$, then $\alpha_i = 0$.

Remark 3.3.1. *The existence of an optimal control for the problem described above is ensured by the convexity of the sets $F(t, \xi) = \{\left(\xi_i + (1 - \alpha_i)\bar{\xi}\right)_{i=1\ldots N},\ \alpha \in [0, 1]^N,\ \sum_i \alpha_i \leq M\}$.[14]*

3.4 Final cost with two agents

For a clearer understanding of the mechanisms taking place, we consider the simple case of two agents in $\mathbb{R}^d$. We consider the sets of controls $\mathcal{U}_M$, where $0 < M \leq 2$. Thus, system (3.7) becomes:

$$\begin{cases} \dot{\xi}_1 = -\xi_1 + (1 - \alpha_1)\, \bar{\xi} \\ \dot{\xi}_2 = -\xi_2 + (1 - \alpha_2)\, \bar{\xi}. \end{cases} \tag{3.22}$$

Computing the difference of the two projected variables will also prove useful:

$$\dot{\xi}_1 - \dot{\xi}_2 = -(\xi_1 - \xi_2) - (\alpha_1 - \alpha_2)\bar{\xi}. \tag{3.23}$$

Three different situations may arise, depending on the value of the constraint on the control. Indeed, two constraints are set: $\alpha_1 + \alpha_2 \leq M$, and $0 \leq \alpha_i \leq 1$ for $i = 1, 2$. We differentiate the cases (a) $0 < M \leq 1$, (b) $1 < M < 2$ and (c) $M = 2$.

3.4.1 Pontryagin's Maximum Principal

Notice that the migration functional can be written as:

$$\mathbb{V} = \frac{1}{2}(\xi_1^2 + \xi_2^2) = \frac{1}{4}\left((\xi_1 + \xi_2)^2 + (\xi_1 - \xi_2)^2\right) = \bar{\xi}^2 + \left(\frac{\xi_1 - \xi_2}{2}\right)^2, \tag{3.24}$$

once again emphasizing the necessary trade-off between two terms: the mean velocity $\bar{\xi}$ and the distance between the agents $|\xi_1 - \xi_2|$. Computing the Hamiltonian of the system gives:

$$H(t, \xi, \lambda, \alpha) = -\bar{\xi}\,(\alpha_1\lambda_1 + \alpha_2\lambda_2) + \frac{\xi_2 - \xi_1}{2}(\lambda_1 - \lambda_2). \tag{3.25}$$

In line with Hyp. 2, two cases are possible: $\xi_1(0) = \xi_2(0)$ or $\xi_1(0) > \xi_2(0)$. The following proposition deals with the first case.

Proposition 3.4.1. *If $\xi_1(0) = \xi_2(0)$, then a control strategy α is optimal if and only if it satisfies $\alpha_1 + \alpha_2 \equiv M$ and $\xi_1(T) = \xi_2(T)$.*

Proof. Consider the control given by $\tilde{\alpha}_1 \equiv \tilde{\alpha}_2 \equiv \frac{M}{2}$. It achieves $[\frac{1}{2}(\xi_1(T) - \xi_2(T))]^2 = 0$ and ensures the maximal decrease of $\bar{\xi}^2$, thus is optimal for the minimization of $\mathbb{V}(T)$ (3.24). Still from (3.24), a control α is optimal if and only if it achieves $[\frac{1}{2}(\xi_1(T) - \xi_2(T))]^2 = 0$, which is equivalent to $\xi_1(T) = \xi_2(T)$, and ensures the maximal decrease of $\bar{\xi}^2$, which is equivalent to $\alpha_1 + \alpha_2 \equiv M$. □

Hence, the case $\xi_1(0) = \xi_2(0)$ is fully understood. In the following, we will deal with more complex cases by assuming:

Hypothesis 4. $\xi_1(0) > \xi_2(0)$.

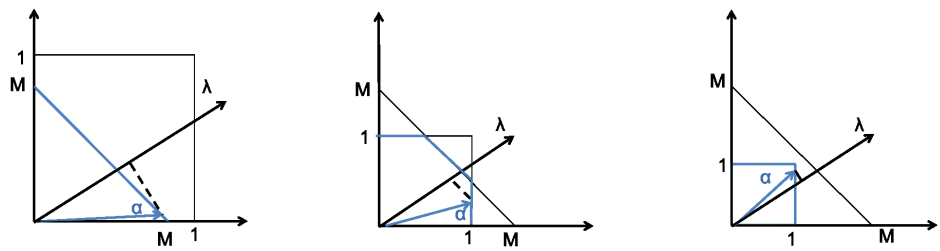
Before studying each case in detail, we give general considerations on the relation between the control α and λ:

(a) If $M \leq 1$, minimizing H (i.e. maximizing $\langle \lambda, \alpha \rangle$) gives (see Fig.3.1a): $(\alpha_1, \alpha_2) = (M, 0)$ if $0 < \lambda_2 < \lambda_1$; $(\alpha_1, \alpha_2) = (M/2, M/2)$ if $0 < \lambda_2 = \lambda_1$; $(\alpha_1, \alpha_2) = (M, 0)$ if $\lambda_2 < 0 < \lambda_1$; $(\alpha_1, \alpha_2) = (0, 0)$ if $\lambda_2 < 0$ and $\lambda_1 < 0$.

(b) If $1 < M < 2$, minimizing H gives (see Fig.3.1b): $(\alpha_1, \alpha_2) = (1, M - 1)$ if $0 < \lambda_2 < \lambda_1$; $(\alpha_1, \alpha_2) = (M/2, M/2)$ if $0 < \lambda_2 = \lambda_1$; $(\alpha_1, \alpha_2) = (1, 0)$ if $\lambda_2 < 0 < \lambda_1$; $(\alpha_1, \alpha_2) = (0, 0)$ if $\lambda_2 < 0$ and $\lambda_1 < 0$.

(c) If $M \geq 2$, minimizing H gives (see Fig.3.1c): $(\alpha_1, \alpha_2) = (1, 1)$ if $0 < \lambda_2 \leq \lambda_1$; $(\alpha_1, \alpha_2) = (1, 0)$ if $\lambda_2 < 0 < \lambda_1$; $(\alpha_1, \alpha_2) = (0, 0)$ if $\lambda_2 < 0$ and $\lambda_1 < 0$.

Notice that in all three cases, if $\lambda_1 = \lambda_2$, then the Pontryagin maximum principle does not give sufficient information since any combination of α_1 and α_2 such that $\alpha_1 + \alpha_2 = M$ minimizes the scalar product $-\langle \lambda, \alpha \rangle$ (see Figure 3.1).



(a) Case $M \leq 1$ (b) Case $1 < M < 2$ (c) Case $M \geq 2$

Figure 3.1: Minimizing $-\langle \lambda, \alpha \rangle$

The dynamics for λ are given by $\dot{\lambda} = -\nabla H = \begin{pmatrix} \frac{1+\alpha_1}{2}\lambda_1 - \frac{1-\alpha_2}{2}\lambda_2 \\ \frac{1+\alpha_2}{2}\lambda_2 - \frac{1-\alpha_1}{2}\lambda_1 \end{pmatrix}$, which allows us to compute the evolution of the difference $\lambda_1 - \lambda_2$:

$$\frac{d}{dt}(\lambda_1 - \lambda_2) = \lambda_1 - \lambda_2. \tag{3.26}$$

The transversality conditions give: $\lambda(T) = \nabla \mathbb{V}(T) = (\xi_1(T), \xi_2(T))^T$. Hence, if the final configuration is such that $\xi_1(T) \neq \xi_2(T)$, i.e. $\lambda_1(T) \neq \lambda_2(T)$, the difference $\lambda_1 - \lambda_2$ increases with time. On the other hand, if $\lambda_1(T) = \lambda_2(T)$, then $\forall t \leq T,\ \lambda_1(t) = \lambda_2(t)$. If the dynamics allow us to drive ξ_1 and ξ_2 together before time T, then $\lambda_1(t) = \lambda_2(t)$ for all t, and the Pontryagin maximum principle does not give sufficient information, as seen above.

3.4.2 Global Strategy

According to equation (3.24), the functional $\mathbb{V}$ can be written as:

$$\mathbb{V} = \bar{\xi}^2 + \frac{(\xi_1 - \xi_2)^2}{4}. \tag{3.27}$$

Minimizing $\mathbb{V}$ requires minimizing $\bar{\xi}$ and $(\xi_1 - \xi_2)^2$ simultaneously. The evolution of $\bar{\xi}$ is given by:

$$\dot{\bar{\xi}} = -\frac{1}{2}(\alpha_1 + \alpha_2)\ \bar{\xi}, \tag{3.28}$$

while that of $(\xi_1 - \xi_2)^2$ is:

$$\frac{d}{dt}\left((\xi_1 - \xi_2)^2\right) = -2(\xi_1 - \xi_2)^2 - 2(\xi_1 - \xi_2)\bar{\xi}(\alpha_1 - \alpha_2). \tag{3.29}$$

Thus, minimizing $\bar{\xi}^2$ (both instantaneously and globally) requires using full control, i.e. setting $\alpha_1 + \alpha_2 = M$. On the other hand, the strategy to minimize $(\xi_1 - \xi_2)^2$ is less clear. It would require both maximizing $\bar{\xi}$ and maximizing the difference $\alpha_1 - \alpha_2$ (assuming that $\xi_1 - \xi_2 \geq 0$), and these conditions might not be compatible.

3.4.3 Case $M = 1$

Theorem 3.4.1.

Let $T > 0$ and let $M = 1$. Furthermore, let $\alpha = (\alpha_1, \alpha_2) \in \mathcal{U}_1$ (see (3.4)) be an optimal control and ξ be the corresponding trajectory of system (3.22). Define $t_0 = 2\ln(\xi_1(0)/\bar{\xi}(0))$. Then

(i) *$T \geq t_0$ if and only if $\xi_1(T) = \xi_2(T)$. In such a case, the control satisfies: $\alpha_1 + \alpha_2 \equiv 1$ (so $\bar{\xi}(t) = \bar{\xi}(0)e^{-t/2}$). For instance, the strategy $(\alpha_1, \alpha_2)(t) = (1, 0)$ for all $t \in [0, t_0[$ and $(\alpha_1, \alpha_2)(t) = (1/2, 1/2)$ for all $t \in [t_0, T]$ is optimal.*

(ii) *If $T < t_0$, then $\alpha(t) = (0, 0)$ for all $t \in [0, t^*[$ and $\alpha(t) = (1, 0)$ for all $t \in [t^*, T]$, where $t^* = 2\ln(\bar{X})$ and $\bar{X} \in [1, e^{T/2}[$ is defined as follows:*

$$\begin{aligned}\bar{X} = \arg\min_{X \in [1, e^{T/2}]} [&\left(\xi_1(0) + \bar{\xi}(0)(X^2 - 1)\right)^2 + \\ &\left(\xi_2(0) + \bar{\xi}(0)(X^2 - 1) + 2\bar{\xi}(0)X(e^{T/2} - X)\right)^2].\end{aligned} \tag{3.30}$$

Proof. Let ξ be an optimal trajectory achieved with optimal control α.

To prove (i), we shall show that the three statements (a) $T \geq t_0$, (b) there exists $t \in [0, T]$ such that $\xi_1(t) = \xi_2(t)$ and (c) $\xi_1(T) = \xi_2(T)$ are equivalent.

Suppose (b) there exists $\tau \in [0, T]$ such that $\xi_1(\tau) = \xi_2(\tau)$. Then necessarily $\xi_1(T) = \xi_2(T)$. Indeed, suppose that $\xi_1(T) \neq \xi_2(T)$. Then any strategy $\tilde{\alpha}$ such that on $[0, \tau]$, $\tilde{\alpha} = \alpha$ and on $]\tau, T]$, $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (\frac{\alpha_1+\alpha_2}{2}, \frac{\alpha_1+\alpha_2}{2})$ achieves: $\bar{\tilde{\xi}}(T) = \bar{\xi}(T)$ and $(\tilde{\xi}_1 - \tilde{\xi}_2)^2(T) = 0 < (\xi_1 - \xi_2)(T)$ (where $\tilde{\xi}$, $\tilde{V}$ denote the trajectory and cost corresponding to $\tilde{\alpha}$), so according to equation (3.27), $\tilde{\mathbb{V}}(T) < \mathbb{V}(T)$ and control strategy α cannot be optimal. Hence, $\xi_1(T) = \xi_2(T)$.

Now suppose (c) $\xi_1(T) = \xi_2(T)$. The transversality condition (3.18) gives $\lambda_1(T) = \lambda_2(T)$ and from Proposition 3.3.1 we get: $\lambda_1(t) = \lambda_2(t)$ for all $t \in [0, T]$. Then, $\dot{\bar{\lambda}} = \sum \alpha_i \lambda_i = (\sum \alpha_i)\bar{\lambda}$. Since $\bar{\xi}(T) > 0$, the transversality condition (3.18) gives: $\bar{\lambda}(T) > 0$, and $\bar{\lambda}(t) = \lambda_1(t) = \lambda_2(t) > 0$

for all $t \in [0,T]$. Therefore, the set I_λ, see (3.21), is not empty, so according to Proposition 3.3.2, the optimal control strategy requires using maximal control strength: $\alpha_1 + \alpha_2 \equiv 1$. According to equation (3.28), this suffices to fully determine $\bar{\xi}(t) = \bar{\xi}(0)\ e^{-t/2}$. Then $\xi_1(t) - \xi_2(t) = e^{-t}\left((\xi_1 - \xi_2)(0) - \bar{\xi}(0)\int_0^t (\alpha_1 - \alpha_2)e^{s/2}ds\right)$, and $\xi_1(t) - \xi_2(t) = 0$ if, and only if, $\int_0^t e^{s/2}(\alpha_1 - \alpha_2)(s)ds = (\xi_1(0) - \xi_2(0))/\bar{\xi}(0)$. Notice that $\min_{(\alpha_1,\alpha_2)\in\mathcal{U}_1}\{t \mid (\xi_1 - \xi_2)(t) = 0\}$ is obtained when $\alpha_1 - \alpha_2$ is maximal, i.e. for $(\alpha_1, \alpha_2) \equiv (1,0)$. With this strategy, $\min_{(\alpha_1,\alpha_2)\in\mathcal{U}_1}\{t \mid (\xi_1 - \xi_2)(t) = 0\} := t_0 = 2\ln(\xi_1(0)/\bar{\xi}(0))$. Hence, we must have: $T \geq t_0$.

Lastly, suppose (a) $T \geq t_0$. Design a strategy $\tilde{\alpha}$ so that for all $t < t_0$, $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (1,0)$ and for all $t \geq t_0$, $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (1/2, 1/2)$. This strategy is optimal since it maximizes the decrease of $\bar{\tilde{\xi}}$, see (3.28), and achieves $(\tilde{\xi}_1 - \tilde{\xi}_2)(T) = 0$, see (3.29). Hence, our optimal strategy α must also satisfy: $\xi_1(T) = \xi_2(T)$ and $\alpha_1 + \alpha_2 \equiv 1$. This proves (b).

We showed that (a), (b) and (c) are equivalent. We thus proved the first part of the proposition: $T \geq t_0$ if and only if $\xi_1(T) = \xi_2(T)$. In this case, it also holds: $\alpha_1 + \alpha_2 \equiv 1$.

If on the other hand, (ii) $T < t_0$, then $\xi_1(t) > \xi_2(t)$ for all $t \in [0,T]$ (since (b) implies (a)). According to condition (3.18) and to Prop. 3.3.1, $\lambda_1(t) > \lambda_2(t)$ for all $t \in [0,T]$ and $\lambda_1(T) > 0$. The evolution of λ_1 is given by: $\dot{\lambda}_1 = \frac{1}{2}(\alpha_1\lambda_1 + \alpha_2\lambda_2) + \lambda_1 - \bar{\lambda} > 0$ since $\lambda_1 > \bar{\lambda}$. Hence, two cases must be distinguished: either $\lambda_1 > 0$ at all time, so the set I_λ^+ is non-empty and full control will be used at all time, or there exists $t^* \in]0,T[$ such that $\lambda_1 < 0$ on $[0,t^*[$, $\lambda_1(t^*) = 0$ and $\lambda_1 > 0$ on $]t^*,T]$, in which case $\alpha = (0,0)$ on $[0,t^*[$ and $\alpha = (1,0)$ on $]t^*,T]$. Knowing this, it is easy to express ξ_1, ξ_2 and $\mathbb{V}$ as functions of t^*:

$$\forall t \in [t^*,T], \begin{cases} \xi_1(t) = e^{-t}(\xi_1(0) + \bar{\xi}(0)(e^{t^*} - 1)) \\ \xi_2(t) = e^{-t}(\xi_2(0) + \bar{\xi}(0)(e^{t^*} - 1) + 2\bar{\xi}(0)e^{t^*/2}(e^{t/2} - e^{t^*/2})) \\ \mathbb{V}(t) = \xi_1^2(t) + \xi_2^2(t) \end{cases} . \tag{3.31}$$

Denoting $X = e^{t^*/2}$, $\mathbb{V}(T)$ can be written as a biquadratic polynomial in X:

$$\begin{aligned} \mathbb{V}(T)(X) = e^{-2T}\Big[&\left(\xi_1(0) + \bar{\xi}(0)(X^2 - 1)\right)^2 + \\ &\left(\xi_2(0) + \bar{\xi}(0)(X^2 - 1) + 2\bar{\xi}(0)X(e^{T/2} - X)\right)^2 \Big]. \end{aligned} \tag{3.32}$$

We look for $\bar{X}$ minimizing $\mathbb{V}(T)(X)$ in the interval $[1, e^{T/2}]$ (so that $t^* \in [0,T]$). Notice that the leading term is $2e^{-2T}\bar{\xi}(0)^2 \cdot X^4$. Hence, there are at most two local minima in the interval $[1, e^{T/2}]$. Furthermore, $\mathbb{V}(T)(1) = e^{-2T}\left[\xi_1(0)^2 + (\xi_2(0) + 2\bar{\xi}(0)(e^{T/2} - 1))^2\right]$ and $\mathbb{V}(T)(e^{T/2}) = e^{-2T}[(\xi_1(0) +$

$\bar{\xi}(0)(e^{T/2}-1))^2 + (\xi_2(0) + \bar{\xi}(0)(e^T-1))^2]$, so $\mathbb{V}(T)(1) < \mathbb{V}(T)(e^{T/2})$, which means that $\bar{X} < e^{T/2}$. If $\bar{X} = 1$, then $t^* = 0$ so it is optimal to act with control $(1,0)$ on the full interval $[0,T]$. If $1 < \bar{X} < e^{T/2}$, then $0 < t^* < T$. The optimal control strategy will require leaving the system to evolve without control on $[0,t^*[$, and acting with control $\alpha = (1,0)$ on $[t^*,T]$. □

Remark 3.4.1. *The existence of an initial "Inactivation" period can be proven also with any number of agents (see Theorem 3.5.3). Numerical simulations with any number of agents (see Section 3.5.2) show that in some cases it is indeed optimal to let the system evolve without control on an initial time interval* $[0,t^*]$*, where* $t^* > 0$.

3.4.4 Case $M < 1$

Generalizing to the case of any $M < 1$, we conduct the same analysis and the optimal control strategy is similar.

Theorem 3.4.2. *Let T>0 and M<1. Let* $\alpha = (\alpha_1, \alpha_2) \in \mathcal{U}_M$ *(see* (3.4)*) be an optimal control and* ξ *be the corresponding trajectory of system* (3.22)*. Define* $t_0 = \frac{2}{2-M} \ln\left(\frac{2-M}{2M}(\xi_1(0) - \xi_2(0))/\bar{\xi}(0) + 1\right)$*. Then*

(i) $T \geq t_0$ *if and only of* $\xi_1(T) = \xi_2(T)$*. In this case, the control satisfies:* $\alpha_1 + \alpha_2 \equiv M$ *(so* $\bar{\xi}(t) = \bar{\xi}(0)e^{-Mt/2}$*).*

(ii) *If* $T < t_0$*, then there exists* $t^* \in [0,T[$ *such that* $\alpha(t) = (0,0)$ *for all* $t \in [0,t^*[$ *and* $\alpha(t) = (1,0)$ *for all* $t \in [t^*,T]$*.*

Remark 3.4.2. *To compute* $t^* \in [0,T[$ *in the case* $T < t_0$*, one can compute* $\mathbb{V}(T)(e^{t^*/2})$ *depending on* $t^* \in [0,T]$ *similarly to the case* $M = 1$*.*

Proof. Let ξ be an optimal trajectory achieved with optimal control $\alpha \in \mathcal{U}_M$. We argue as in the case $M = 1$.

To prove (i), first suppose that there exists $\tau \in [0,T]$ such that $\xi_1(\tau) = \xi_2(\tau)$. Then, as in the case $M = 1$, necessarily it holds $\xi_1(T) = \xi_2(T)$ and any strategy achieving $\xi_1(T) = \xi_2(T)$ while using maximum control $\alpha_1 + \alpha_2 \equiv M$ is optimal. Then $\xi_1(t) - \xi_2(t) = 0 \Leftrightarrow \int_0^t e^{\frac{2-M}{2}s}(\alpha_1 - \alpha_2)(s)ds = (\xi_1(0) - \xi_2(0))/\bar{\xi}(0)$. Hence, $\min_{\alpha \in \mathcal{U}_M}\{t \mid (\xi_1 - \xi_2)(t) = 0\}$ is obtained when $\alpha_1 - \alpha_2$ is maximal, i.e. for $(\alpha_1, \alpha_2) \equiv (M, 0)$. With this strategy, $\min_{\alpha \in \mathcal{U}_M}\{t \mid (\xi_1 - \xi_2)(t) = 0\} = t_0$ as defined above. Hence, if there exists $\tau \leq T$ such that $\xi_1(\tau) = \xi_2(\tau)$, then $T \geq t_0$. This proves the first implication of the proposition: if $\xi_1(T) = \xi_2(T)$, then $T \geq t_0$.

Conversely, if $T \geq t_0$, then the strategy $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (M, 0)$ on $[0, t_0[$ and $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (\frac{M}{2}, \frac{M}{2})$ on $[t_0, T]$ is optimal since it minimizes $\bar{\tilde{\xi}}(T)$ and achieves $\tilde{\xi}_1(T) = \tilde{\xi}_2(T)$. Hence, if α is optimal, it must also satisfy $\alpha_1 + \alpha_2 \equiv M$ and $\xi_1(T) = \xi_2(T)$, which proves the second implication.

Now assume (ii) $T < t_0$. From (i) we get: $\xi_1(t) > \xi_2(t)$ for all $t \in [0, T]$. One can then argue as in the case $M = 1$. According to Pontryagin's Maximum Principle, $\alpha_2 \equiv 0$ and two cases have to be distinguished: either $\lambda_1 > 0$ at all time, so the set I_λ^+ (see (3.21)) is non-empty and full control will be used at all time, or there exists $t^* \in]0, T[$ such that $\lambda_1 < 0$ on $[0, t^*[$, $\lambda_1(t^*) = 0$ and $\lambda_1 > 0$ on $]t^*, T]$, in which case $\alpha = (0, 0)$ on $[0, t^*[$ and $\alpha = (M, 0)$ on $]t^*, T]$.

□

Remark 3.4.3. *Notice that in the limit case $M \to 1$ of Theorem 3.4.2, one finds the same expression for t_0 as in Theorem* 3.4.1.

3.4.5 Case $M = 2$

In order to determine the optimal strategy, let us first study the evolution of the covectors λ. From $\xi_1(T) \geq \bar{\xi}(T) > 0$ (see Prop. 3.1.1 and Hyp. 3) and the transversality condition (3.18), we get $\lambda_1(T) > 0$.

Proposition 3.4.2. *Let $M = 2$ and λ_1 and λ_2 be the covectors corresponding to an optimal control strategy for the system* (3.22)*. Then they satisfy the following properties:*

(i) *If $\lambda_2(T) > 0$, then $\lambda_1(t) > 0$ and $\lambda_2(t) > 0$ for all $t \in [0, T]$.*

(ii) *If $\lambda_2(T) = 0$, then $\lambda_1(t) > 0$ and $\lambda_2(t) = 0$ for all $t \in [0, T]$.*

(iii) *If $\lambda_2(T) < 0$, then $\lambda_2(t) < 0$ for all $t \in [0, T]$.*

Proof.

(i) Let $\lambda_2(T) > 0$. Suppose that there exists $\tau \in [0, T[$ such that $\lambda_2(\tau) = 0$ and $\lambda_2(t) > 0$ for all $t \in]\tau, T]$. Then since $\lambda_1 \geq \lambda_2 > 0$ on $]\tau, T]$, according to Pontryagin's maximum principle (see Section 3.4.1), $(\alpha_1, \alpha_2) \equiv (1, 1)$ on $]\tau, T]$, which gives the following evolutions: $\dot{\lambda}_1 = \lambda_1$ and $\dot{\lambda}_2 = \lambda_2$. Hence, $\lambda_2(\tau) = \lambda_2(T)e^{\tau - T} > 0$, which contradicts the definition of τ. Therefore, $\lambda_2(t) > 0$ for all $t \in [0, T]$, and by (3.20), $\lambda_1(t) > 0$.

(ii) Let $\lambda_2(T) = 0$. Let $\tau := \inf_{[0,T]}\{\bar{t} \in [0, T] \text{ s.t. } \lambda_2(t) = 0 \text{ for all } t > \bar{t}\}$ and suppose that $\tau > 0$. By definition of τ, $\lambda_2(\tau) = 0$. Since $\lambda_1(t) > \lambda_2(t)$ for all t (see Prop. 3.3.1), there exists an interval $[\tau - \delta, \tau[$ on which $\lambda_1 > 0$ and either $\lambda_2 > 0$ or $\lambda_2 < 0$. If $\lambda_2(t) > 0$ for all $t \in [\tau - \delta, \tau[$, then

according to Pontryagin's maximum principle (Section 3.4.1), the control satisfies $\alpha_1(t) = \alpha_2(t) = 1$, which gives: $\dot{\lambda}_2(t) = \lambda_2(t) > 0$. So $\lambda_2(\tau) > 0$, which contradicts the definition of τ. If on the other hand $\lambda_2(t) < 0$ for all $t \in [\tau - \delta, \tau[$, then $\alpha_2(t) = 0$ and $\dot{\lambda}_2(t) = \frac{1}{2}\lambda_2(t) < 0$, which is impossible since it implies $\lambda_2(\tau) < 0$. Hence, $\tau = 0$. Furthermore, since $\lambda_1(T) > 0$ and $\lambda_2 \equiv 0$, then $\dot{\lambda}_1 = \lambda_1$ in a neighborhood of T, which ensures that $\lambda_1(t) > 0$ for all $t \in [0, T]$ (by the same reasoning as in (i)).
(iii) Let $\lambda_2(T) < 0$. Define $\tau := \inf_{[0,T]}\{\bar{t} \in [0, T] \text{ s.t. } \lambda_1(t) > 0 \text{ and } \lambda_2(t) < 0 \text{ for all } t > \bar{t}\}$. Then on $]\tau, T]$, as seen in Section 3.4.1, $\alpha_1 \equiv 1$ and $\alpha_2 \equiv 0$, which gives: $\lambda_2(t) = \lambda_2(\tau)e^{T-\tau}$. Since $\lambda_2(T) < 0$, it follows that $\lambda_2(\tau) < 0$. Hence, either $\tau = 0$ or $\lambda_1(\tau) = 0$. Notice that since $\lambda_1(t) > \lambda_2(t)$ for all t, λ_1 is strictly increasing (see (3.16)). Then the former case implies that $\lambda_2(t) < 0$ for all $t \in [0, T]$. In the latter case, we get that $\lambda_2(t) < 0$ for all $t \leq \tau$. □

This information about the covectors allows us to solve the optimization problem based on the initial conditions and the final time. Recall from Proposition 3.1.1 that $\xi_1(0) > 0$.

Theorem 3.4.3. *Let M=2. Let $(\alpha_1, \alpha_2) \in \mathcal{U}_2$ be an optimal control strategy and ξ be the corresponding trajectory for system* (3.22)*. Define $t_0 = 2\ln\big(\xi_1(0)/(2\bar{\xi}(0))\big)$.*

(i) *If $\xi_2(0) > 0$, then $(\alpha_1, \alpha_2) \equiv (1, 1)$.*

(ii) *If $\xi_2(0) \leq 0$ and $T \geq t_0$, then $\xi_2(T) = 0$ and $\alpha_1 \equiv 1$. For instance the strategy $(\alpha_1, \alpha_2) = (1, 0)$ for all $t \in [0, t_0[$ and $(\alpha_1, \alpha_2) = (1, 1)$ for all $t \in [t_0, T]$ is optimal. Furthermore, if there exists $\bar{t} \in [0, T[$ such that $\xi_2(\bar{t}) = 0$, then $\xi_2(t) = 0$ for all $t \in [\bar{t}, T]$.*

(iii) *If $\xi_2(0) \leq 0$ and $T < t_0$, then there exists $t^* \in [0, T[$ such that $\alpha(t) = (0, 0)$ for all $t \in [0, t^*[$ and $\alpha(t) = (1, 0)$ for all $t \in [t^*, T]$.*

Proof. Let (α_1, α_2) be an optimal control strategy and ξ be the corresponding trajectory.
(i) Let $\xi_2(0) > 0$. According to Prop. 3.1.1, for all $t \in [0, T]$ it holds $\xi_1(t) > 0$ and $\xi_2(t) > 0$. Then $\lambda_1(T) > 0$ and $\lambda_2(T) > 0$. From Prop. 3.4.2 it follows that $\lambda_1(t) > 0$ and $\lambda_2(t) > 0$ for all $t \in [0, T]$. According to the PMP (see Section 3.4.1), maximal control has to be used at all time, i.e. $(\alpha_1, \alpha_2)(t) = (1, 1)$ for all $t \in [0, T]$.

For cases (ii) and (iii), let $\xi_2(0) \leq 0$. By Prop. 3.1.1 it holds $\xi_1(t) > 0$ for all $t \in [0, T]$. Suppose that $\xi_2(T) > 0$. Then from Prop. 3.4.2 we get $\lambda_1(t) \geq \lambda_2(t) > 0$ for all $t \in [0, T]$, so $(\alpha_1, \alpha_2) \equiv (1, 1)$. But with this strategy $\dot{\xi}_2 = -\xi_2$, so $\xi_2(t) = \xi_2(0)e^{-t} \leq 0$ for all $t \in [0, T]$, which contradicts $\xi_2(T) > 0$. Hence $\xi_2(T) \leq 0$.
(ii) First assume that $T \geq t_0$. Let us show that $\xi_2(T) = 0$ and $\alpha_1 \equiv 1$. Such a strategy exists, since for instance the control $(\beta_1, \beta_2)(t) = (1, 0)$ for $t \in [0, t_0[$ and $(\beta_1, \beta_2)(t) = (1, 1)$ for $t \in [t_0, T]$

achieves $\xi_2^\beta(t) = 0$ for all $t \in [t_0, T]$ (where ξ^β denotes the trajectory corresponding to the control strategy β) – by direct computation of (3.22). Suppose that $\xi_2(T) < 0$. Then α cannot be optimal since the control strategy β achieves the minimum of $\xi_1^\beta(T)^2$, see (3.22), and of $\xi_2^\beta(T)^2$ and therefore the minimum of $\mathbb{V}(T) = \xi_1^\beta(T)^2 + \xi_2^\beta(T)^2$. Hence α must satisfy $\alpha_1 \equiv 1$ and $\xi_2(T) = 0$ in order to perform as well as β. Obviously, all strategies that achieve $\xi_2(T) = 0$ with $\alpha_1 \equiv 1$ achieve the same final positions (see (3.22)) and thus have the same $\mathbb{V}(T)$. Furthermore, if there exists a $\hat{t} < T$ such that $\xi_2(\hat{t}) = 0$, then $\xi_2(t) = 0$ for all $t \in [\hat{t}, T]$: if $\xi_2(\bar{t}) = 0$, then $\dot{\xi}_2(\bar{t}) = (1 - \alpha_2)\bar{\xi}(\bar{t}) \geq 0$ and therefore ξ_2 cannot become negative, once it reaches 0. On the other hand, if $\xi_2(t) > 0$, then $\xi_2(T) > 0$ by Prop. 3.1.1.

(iii) Assume now that $T < t_0$. Firstly, we show that an optimal strategy (by PMP) always achieves $\xi_2(T) < 0$. We argue by contradiction: Assume that $\xi_2(T) = 0$. Then $\lambda_1(T) > 0$ and $\lambda_2(T) = 0$ and, according to Proposition 3.4.2, it follows that $\lambda_1(t) > \lambda_2(t) = 0$ for all $t \in [0, T]$. According to the PMP, $\alpha_1 \equiv 1$. Then the growth of ξ_2 is maximal if, and only if, $\alpha_2 \equiv 0$ since in this case $\bar{\xi}$ is maximal. But with this strategy ξ_2 cannot reach 0 before t_0 – by direct computation of (3.22). Therefore $\xi_2(T) < 0$, so $\lambda_2(T) < 0$ and $\lambda_2(t) < 0$ for all $t \in [0, T]$ by Prop. 3.4.2. Hence we are in the same situation as in the case $M = 1$ and $M < 1$. Two cases are possible: either $\lambda_1 > 0$ at all time, so the set I_λ^+ is non-empty and full control on ξ_1 is used at all time, or there exists $t^* \in]0, T[$ such that $\lambda_1 < 0$ on $[0, t^*[$, $\lambda_1(t^*) = 0$ and $\lambda_1 > 0$ on $]t^*, T]$, in which case $\alpha = (0, 0)$ on $[0, t^*[$ and $\alpha = (1, 0)$ on $]t^*, T]$. □

Remark 3.4.4. *To compute $t^* \in [0, T[$ in the case $\xi_1(0) > -\xi_2(0) > 0$ and $T < t_0$, one can compute $\mathbb{V}(T)(X)$ depending on $t^* \in [0, T[$ similarly to the case $M = 1$.*

3.4.6 Case $1 < M < 2$

As in the case $M = 2$, we state the following properties concerning the covectors λ.

Proposition 3.4.3. *Let $M \in]1, 2[$ and λ_1 and λ_2 be the covectors corresponding to an optimal control strategy for the system* (3.22)*. They satisfy the following properties:*

(i) *If $\lambda_2(T) > 0$, then $\lambda_1(t) > 0$ and $\lambda_2(t) > 0$ for all $t \in [0, T]$.*

(ii) *If $\lambda_2(T) = 0$, then $\lambda_1(t) > 0$ and $\lambda_2(t) = 0$ for all $t \in [0, T]$.*

(iii) *If $\lambda_2(T) < 0$, then $\lambda_2(t) < 0$ for all $t \in [0, T]$.*

Proof. The proof is very similar to that of Prop 3.4.2.

(i) Let $\lambda_2(T) > 0$. Suppose that there exists $\tau \in [0, T[$ such that $\lambda_2(\tau) = 0$ and $\lambda_2(t) > 0$ for all

$t \in]\tau, T]$. Then if $\lambda_1 > \lambda_2 > 0$ on $]\tau, T]$, according to Pontryagin's maximum principle (see Section 3.4.1), $(\alpha_1, \alpha_2) \equiv (1, M-1)$ on $]\tau, T]$, which gives $\dot{\lambda}_2 = \frac{M}{2}\lambda_2$. If $\lambda_1 = \lambda_2 > 0$ on $]\tau, T]$, then $\alpha_1 + \alpha_2 \equiv M$ (see Figure 3.1b), which also gives $\dot{\lambda}_2 = \frac{M}{2}\lambda_2$. Hence, $\lambda_2(\tau) = \lambda_2(T)e^{\frac{M}{2}(\tau - T)} > 0$, which contradicts the definition of τ.

For (ii) and (iii) we reason the same way as in the proof of Proposition 3.4.2. □

As in the previous sections, this allows us to solve the optimal control problem by distinguishing cases based on the initial conditions and the final time. The case $\xi_2(0) < 0$ is illustrated in Figure 3.2.

Theorem 3.4.4.

Let $M \in]1, 2[$. Let $\alpha \in \mathcal{U}_M$ be an optimal control strategy and ξ be the corresponding trajectory. Define $t_0 \leq t_1 \leq t_2$ as: $t_0 = 2\ln\left(\frac{\xi_1(0)}{2\xi(0)}\right)$, $t_1 = \frac{2}{2-M}\ln\left(\frac{\xi_1(0)}{2\xi(0)}\right)$ and $t_2 = \frac{2}{2-M}\ln(\frac{\xi_1(0)}{\xi(0)})$.

If $\xi_2(0) > 0$, two subcases are to be distinguished:

(i) *If $T < t_2$, then $(\alpha_1, \alpha_2) \equiv (1, M-1)$ and $0 < \xi_2(T) < \xi_1(T)$.*

(ii) *If $T \geq t_2$, $\xi_1(T) = \xi_2(T)$ and $\alpha_1 + \alpha_2 = M$.*

In the case $\xi_2(0) < 0$, four subcases appear:

(iii) *If $T < t_0$, then $\xi_2(t) < 0$ and there exists $t^* \in [0, T[$ such that $(\alpha_1, \alpha_2)(t) = (0, 0)$ for all $t \in [0, t^*]$ and $(\alpha_1, \alpha_2)(t) = (1, 0)$ for all $t \in]t^*, T]$.*

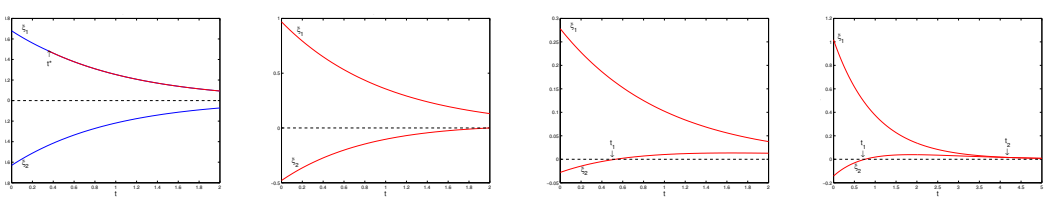
(iv) *If $t_0 \leq T \leq t_1$, then $\alpha_1 \equiv 1$ and $\xi_2(T) = 0$.*

(v) *If $t_1 < T < t_2$, then $(\alpha_1, \alpha_2) \equiv (1, M-1)$ and $0 < \xi_2(T) < \xi_1(T)$.*

(vi) *If $t_2 \leq T$, then $\alpha_1 + \alpha_2 \equiv M$ and $\xi_1(T) = \xi_2(T)$.*

Remark 3.4.5. *Notice that if $\xi_1(0) = \xi_2(0)$, then $t_2 = 0$.*

Remark 3.4.6. *In the limit case $M \to 1$, the times t_0 and t_1 are equal, which is in line with Theorem 3.4.1. In the limit case $M \to 2$, t_1 and t_2 are undefined, in line with Theorem 3.4.3.*



(a) $T < t_0$ (b) $t_0 \leq T \leq t_1$ (c) $t_1 < T < t_2$ (d) $t_2 \leq T$

Figure 3.2: Control strategies in the case $\xi_2(0) < 0$ (controlled agents in red, uncontrolled ones in blue)

Proof. First, let $\xi_1(0) \geq \xi_2(0) > 0$. According to Prop. 3.1.1, $\xi_1(t) > 0$ and $\xi_2(t) > 0$ for all $t \in [0, T]$. The transversality condition gives $\lambda_1(T) > 0$ and $\lambda_2(T) > 0$, and according to Prop. 3.4.3, $\lambda_1(t) > 0$ and $\lambda_2(t) > 0$ for all $t \leq T$. According to Pontryagin's maximum principle (see Section 3.4.1), the global strategy requires setting $\alpha_1 + \alpha_2 \equiv M$. In this case, $\bar{\xi}(t) = \bar{\xi}(0)\exp(-\frac{M}{2}t)$ does not depend on the choice of α_1 and α_2. Minimizing $\mathbb{V}$ (3.27) therefore amounts to minimizing $(\xi_1 - \xi_2)^2$.

(i) If $T \geq t_2$, we will show that in addition to satisfying $\alpha_1 + \alpha_2 \equiv M$, the optimal control α must achieve $\xi_1(T) = \xi_2(T)$. Such a control strategy exists, since for instance (as one can see by direct computation of (3.22)) the control $(\beta_1, \beta_2)(t) = (1, M-1)$ for all $t \in [0, t_2[$ and $(\beta_1, \beta_2)(t) = (M/2, M/2)$ for all $t \in [t_2, T]$ achieves $\xi_1^\beta(t) = \xi_2^\beta(t)$ for all $t \in [t_2, T]$, where ξ^β denotes the corresponding trajectory. Notice that β minimizes $\bar{\xi}(T)$ by using the full strength M of the control at all time (see (3.28)), and minimizes $(\xi_1 - \xi_2)^2(T)$, so it minimizes $\mathbb{V}(T)$ (see (3.24)). Hence, in order to be optimal, α must satisfy $\xi_1(T) = \xi_2(T)$ as well as $\alpha_1 + \alpha_2 \equiv M$.

(ii) If $T < t_2$, we will show that $(\alpha_1, \alpha_2) \equiv (1, M-1)$ and that ξ_1 and ξ_2 cannot be brought together (i.e. $\xi_1(T) > \xi_2(T)$). Indeed, knowing that $\alpha_1 + \alpha_2 \equiv M$, one can use (3.23) to compute: $(\xi_1 - \xi_2)(t) = e^{-t}\Big((\xi_1 - \xi_2)(0) - \int_0^t (\alpha_1 - \alpha_2)(s)\bar{\xi}(s)e^s ds\Big)$. Since $\bar{\xi}$ is fully determined, $t_{\min} := \min_{\alpha \in \mathcal{U}_M, \alpha_1 + \alpha_2 \equiv M}\{t \in [0, T] \text{ s.t. } (\xi_1 - \xi_2)(t) = 0\}$ is achieved by maximizing $(\alpha_1 - \alpha_2)$, which gives: $(\alpha_1, \alpha_2) \equiv (1, M-1)$. As seen previously, by direct computation of (3.22), $t_{\min} = t_2$ as defined above. Hence, if $T < t_2$, necessarily $\xi_1(T) > \xi_2(T)$. Then $\lambda_1(T) > \lambda_2(T)$ and according to Prop. 3.4.3, and to Prop. 3.3.1, $\lambda_1(t) > \lambda_2(t) > 0$ for all t. According to the PMP (see 3.4.1), the optimal strategy is $(\alpha_1, \alpha_2) \equiv (1, M-1)$.

Now let $\xi_1(0) > 0$, $\xi_2(0) < 0$ and $\bar{\xi}(0) > 0$. We then distinguish four subcases.
Firstly, let us prove that if $\xi_2(T) > 0$, then necessarily $T > t_1$. Indeed, if $0 < \xi_2(T) \leq \xi_1(T)$, then $0 < \lambda_2(T) \leq \lambda_1(T)$, and according to Proposition 3.4.3, $0 < \lambda_2(t) \leq \lambda_1(t)$ for all $t \in [0, T]$. According to the PMP (see Section 3.4.1), $\alpha_1 + \alpha_2 \equiv M$. Hence $\bar{\xi}(t) = \bar{\xi}(0)e^{-Mt/2}$, and $\xi_2(t) = e^{-t}(\xi_2(0) + \bar{\xi}(0)\int_0^t (1 - \alpha_2)e^{\frac{2-M}{2}s}ds)$. The minimum time $t_{\min}$ needed to achieve $\xi_2(t_{\min}) > 0$ is achieved for $(\alpha_1, \alpha_2) \equiv (1, M-1)$, which, after computation, gives $t_{\min} = t_1$ as defined above. Hence, if $\xi_2(T) \geq 0$, then $T > t_1$.

(iii) Let $T \geq t_2$. Let us prove that $\alpha_1 + \alpha_2 \equiv M$ and $\xi_1(T) = \xi_2(T)$. Such a control strategy exists. Indeed, take for example $(\beta_1, \beta_2)(t) = (1, M-1)$ on $[0, t_2]$ and $(\beta_1, \beta_2)(t) = (M/2, M/2)$ on $]t_2, T]$. Then, by direct computation of (3.22), $\xi_1^\beta(t) = \xi_2^\beta(t)$ for all $t \in [t_2, T]$ (where ξ^β denotes the trajectory corresponding to the control β). Furthermore, β is optimal since it minimizes $\bar{\xi^\beta}$ by using full control at all time and achieves $(\xi_1^\beta - \xi_2^\beta)^2(T) = 0$ (see (3.24)). In order to perform optimally,

the control α must also satisfy $\alpha_1 + \alpha_2 \equiv M$ and $\xi_1(T) = \xi_2(T)$.

(iv) Let $T < t_0$. Since $t_0 < t_1$, then as proved above, $\xi_2(T) \leq 0$. Suppose that $\xi_2(T) = 0$. Then $\lambda_1(T) > \lambda_2(T) = 0$ and according to Proposition 3.4.3, $\lambda_1(t) > \lambda_2(t) = 0$ for all time t. Hence, $\alpha_1 \equiv 1$ (see Section 3.4.1). Then $\min_{\alpha_2}\{t \in [0,T] \text{ s.t. } \xi_2(t) = 0\} = t_0$ as defined above (obtained for $\alpha_2 \equiv 0$). This contradicts the condition on T. Hence, if $T < t_0$, then $\xi_2(T) < 0$ and according to Proposition 3.4.3 and Section 3.4.1, $\lambda_2 < 0$ so $\alpha_2 \equiv 0$. However, there is no information on λ_1 other than $\dot{\lambda}_1 = \alpha_1/2\lambda_1 + \bar{\lambda} - \lambda_1 \geq 0$ and $\lambda_1(\tau) = 0$ implies $\dot{\lambda}_1(\tau) > 0$. Hence, as in the previous sections, there exists $t^* \in [0,T[$ such that $\lambda_1 < 0$ on $[0,t^*[$, $\lambda_1(t^*) = 0$ and $\lambda_1 > 0$ on $]t^*,T]$. This implies that $(\alpha_1,\alpha_2) = (0,0)$ on $[0,t^*[$ and $(\alpha_1,\alpha_2) = (1,0)$ on $[t^*,T[$.

(v) Let $t_0 \leq T \leq t_1$. We shall prove that $\xi_2(T) = 0$ and that $\alpha_1 \equiv 1$. As seen previously, if $T \leq t_1$, then $\xi_2(T) \leq 0$. Suppose that $\xi_2(T) < 0$. Then $\lambda_1(T) > 0$ and $\lambda_2(T) < 0$ which according to Proposition 3.4.3 gives $\lambda_2(t) < 0$ for all t, and according to the PMP (see Section 3.4.1), $\alpha_2 \equiv 0$. Then $\xi_2(t) = e^{-t}(\xi_2(0) + \bar{\xi}(0)\int_0^T e^{-\int_0^s \frac{1}{2}\alpha_1(r)dr}e^s ds)$. Thus $t_{\sup} := \sup_{\alpha_1}\{\tau \in [0,T] \text{ s.t. } \xi_2(t) < 0 \text{ for all } t \in [0,\tau[\}$ is obtained for $\alpha_1 \equiv 1$ and by direct computation, $t_{\sup} = t_0$. Since $T \geq t_0$, there exists $\tau \leq T$ such that $\xi_2(\tau) = 0$. However, by Proposition 3.1.1, once $\xi_2 = 0$ it cannot become negative again, which contradicts $\xi(T) < 0$. Therefore, $\xi_2(T) = 0$, and according to Proposition 3.4.3 and the PMP (Section 3.4.1), $\lambda_1(t) > 0$ for all $t \in [0,T]$ so $\alpha_1 \equiv 1$. Furthermore, if $\xi_2(\tau) = 0$, then $\dot{\xi}_2(\tau) = (1 - \alpha_2(\tau))\bar{\xi}(\tau) > 0$ since $\alpha_2 = M - \alpha_1 = M - 1 < 1$. According to Proposition 3.1.1, once ξ_2 becomes positive it cannot become zero again. Hence we must have $\xi_2(t) < 0$ for all $t < T$ and $\xi_2(T) = 0$.

(vi) Let $t_1 < T < t_2$. As in the previous case, since $T \geq t_0$, one must have: $\xi_2(T) \geq 0$. Suppose that $\xi_2(T) = 0$. Then according to Proposition 3.4.3 and the PMP, $\alpha_1 \equiv 1$ and $\xi_2(t) = e^{-t}(\xi_2(0) + \bar{\xi}(0)\int_0^T (1-\alpha_2)(s)e^{-\int_0^s \frac{1}{2}(1+\alpha_2)(r)dr}e^s ds)$. Then the minimum of $\xi_2(T)$ is obtained for $\alpha_2 \equiv M - 1$, so

$$\begin{aligned} \xi_2(T) &\geq e^{-T}(\xi_2(0) + \bar{\xi}(0)\int_0^T (2-M)e^{-\frac{1}{2}Ms}e^s ds) \\ &> e^{-T}(\xi_2(0) + \bar{\xi}(0)(e^{\frac{2-M}{2}t_1} - 1)) > 0 \end{aligned} \tag{3.33}$$

by definition of t_1. This contradicts $\xi_2(T) = 0$, so necessarily $\xi_2(T) > 0$. Then $\lambda_1(t) > 0$ and $\lambda_2(t) > 0$ for all t, which implies that $\alpha_1 + \alpha_2 \equiv M$. In this case we prove as in case (ii) that $\xi_1(T) > \xi_2(T)$, which implies $(\alpha_1,\alpha_2) \equiv (1, M-1)$. $\square$

3.5 Final cost with any number of agents and control bounded by M=1

3.5.1 Theroretical Analysis

In this section, we address the optimal control problem of minimizing $\mathbb{V}(T)$ with any number of agents, setting the upper bound $M = 1$ on the strength of the control, i.e. $\sum_{i=1}^{N} \alpha_i \leq 1$. We define the set of such controls:

$$\mathcal{U} = \Big\{\alpha : [0,T] \to [0,1]^N \Big| \ \alpha \text{ measurable, s.t. for all } t \in [0,T] \ \sum_{i=1}^{N} \alpha_i(t) \leq 1\Big\}. \tag{3.34}$$

We remind the equations governing the evolution of ξ_i and $\bar{\xi}$ for $i \in \{1, ..., N\}$:

$$\dot{\xi}_i = -\xi_i + (1-\alpha_i)\bar{\xi} \qquad \text{and} \qquad \dot{\bar{\xi}} = -(\sum_i \alpha_i)\ \bar{\xi}. \tag{3.35}$$

As before, we aim to minimize the migration functional $\mathbb{V} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^2$ over the space $\mathcal{U}$ at final time:

Problem 1. *Find* $\arg\min_{\alpha \in \mathcal{U}} \mathbb{V}(T)$.

Let us consider the restricted set of full-strength controls $\mathcal{U}_{FS} \subset \mathcal{U}$:

$$\mathcal{U}_{FS} = \Big\{\alpha : [0,T] \to [0,1]^N \Big| \ \alpha \text{ measurable, s.t. for all } t, \ \sum_{i=1}^{N} \alpha_i(t) = 1\Big\}. \tag{3.36}$$

We also introduce the set of optimal controls $\mathcal{U}_{\text{opt}}$:

$$\mathcal{U}_{\text{opt}} = \Big\{\alpha \in \mathcal{U} \ \text{ s.t. } \mathbb{V}_\alpha = \min_{\beta \in \mathcal{U}} \mathbb{V}_\beta\Big\}. \tag{3.37}$$

A question then arises naturally: are there optimal controls among full-strength controls? In other words, we study the intersection $\mathcal{U}_{\text{FS}} \cap \mathcal{U}_{\text{opt}}$. To answer this, we first look for an optimal control strategy among the restricted set of controls $\mathcal{U}_{\text{FS}}$, i.e. we consider the problem:

Problem 2. *Find* $\arg\min_{\alpha \in \mathcal{U}_{FS}} \mathbb{V}(T)$.

Introducing the partial mean $\bar{\xi}_{1,l} = \frac{1}{l}\sum_{i=1}^{l} \xi_i$, we design the following optimal control strategy to solve Problem 2.

Theorem 3.5.1 (Full-control strategy)**.**

Let $T > 0$. The strategy designed in Prop 3.2.1 to decrease $\dot{\mathbb{V}}$ instantaneously is an optimal control strategy for Problem 2. It can be explicitly described as follows:

Define $t_1 = 0$ and for $l \in \{2, ..., N\}$, $t_l = \frac{N}{N-1} \ln\left((l-1)\frac{N-1}{N}\frac{\bar{\xi}_{1,l-1}(0)-\xi_l(0)}{\bar{\xi}(0)} + 1\right)$.

If there exists $l \in \{1, ..., N-1\}$ such that $T \in [t_l, t_{l+1}[$, then any strategy satisfying: $\xi_i(T) = \bar{\xi}_{1,l}(T)$ for every $i \in \{1, ..., l\}$, $\sum_{i=1}^{l} \alpha_i \equiv 1$ and $\alpha_i \equiv 0$ for every $i \in \{l+1, ..., N\}$ is optimal.

If $T \geq t_N$, then any strategy satisfying $\xi_i(T) = \bar{\xi}(T)$ for all $i \in \{1, ..., N\}$ and $\sum_{i=1}^{N} \alpha_i \equiv 1$ is optimal.

For instance, if $T \in [t_l, t_{l+1}[$, one optimal strategy would consist in defining the following piecewise constant control:

$$\forall k \leq l,\ \forall t \in [t_k, t_{k+1}[,\ \begin{cases} \alpha_i(t) = \frac{1}{k} \text{ if } i \leq k \\ \alpha_i(t) = 0 \text{ if } i > k. \end{cases} \tag{3.38}$$

Proof. Let us first show that if $T \geq t_l$, then the optimal control strategy for Problem 2 must achieve $\xi_i(T) = \bar{\xi}_{1,l}(T)$ for all $i \in \{1, ..., l\}$, reasoning by contradiction.

Suppose that there exists $k \in \{1, ..., l\}$ such that $\xi_k(T) \neq \bar{\xi}_{1,l}(T)$. Using Hyp. 3, we can suppose that there exists $m < l$ such that for every $i \in \{1, ..., m\}$, $\xi_i(T) = \bar{\xi}_{1,m}(T)$, and for every $i \in \{1, ..., m\}$ and $j \in \{m+1, ..., N\}$, $\xi_j(T) < \xi_i(T)$.

Let $j \in \{m+1, ..., l\}$. The transversality condition (3.18) gives: for every $i \in \{1, ..., m\}$, $\lambda_j(T) < \lambda_i(T)$. According to Proposition 3.3.1, for all $t \in [0, T]$, for every $i \in \{1, ..., m\}$, $\lambda_j(t) < \lambda_i(t)$. According to the PMP, as seen in Section 3.3, only the biggest covectors are controlled, and since $\alpha \in \mathcal{U}_{\text{FS}}$, with maximum control. So $\sum_{i=1}^{m} \alpha_i \equiv 1$ and $\alpha_j \equiv 0$. The evolutions of ξ_j and $\bar{\xi}_{1,m}$ are then given by:

$$\begin{cases} \dot{\xi}_j = -\xi_j + \bar{\xi} \\ \dot{\bar{\xi}}_{1,m} = -\bar{\xi}_{1,m} + \frac{m-1}{m}\bar{\xi}. \end{cases} \tag{3.39}$$

Since $\sum_{i=1}^{N} \alpha_i \equiv 1$, the evolution of the mean is given by $\dot{\bar{\xi}} = -\frac{1}{N}\bar{\xi}$, and we can compute $\bar{\xi} = \bar{\xi}(0)e^{-t/N}$, which in turn allows us to solve:

$$\forall t \in [0, T], \begin{cases} \xi_j(t) = e^{-t}\left(\xi_j(0) + \frac{N}{N-1}\bar{\xi}(0)(e^{\frac{N-1}{N}t} - 1)\right) \\ \bar{\xi}_{1,m}(t) = e^{-t}\left(\bar{\xi}_{1,m}(0) + \frac{m-1}{m}\frac{N}{N-1}\bar{\xi}(0)(e^{\frac{N-1}{N}t} - 1)\right). \end{cases} \tag{3.40}$$

We get:

$$(\bar{\xi}_{1,m} - \xi_j)(T) = e^{-T}\left(\bar{\xi}_{1,m}(0) - \xi_j(0) - \frac{1}{m}\frac{N}{N-1}\bar{\xi}(0)(e^{\frac{N-1}{N}T} - 1)\right). \tag{3.41}$$

We made the hypothesis that $T \geq t_l = \frac{N}{N-1} \ln\left((l-1)\frac{N-1}{N}\frac{\bar{\xi}_{1,l-1}(0)-\xi_l(0)}{\bar{\xi}(0)}+1\right)$. Hence,

$$\begin{aligned}
(\bar{\xi}_{1,m}-\xi_j)(T) &\leq e^{-T}\left(\bar{\xi}_{1,m}(0)-\xi_j(0)-\frac{1}{m}(l-1)(\bar{\xi}_{1,l-1}(0)-\xi_l(0))\right)\\
&= \frac{1}{m}e^{-T}\left[m\bar{\xi}_{1,m}(0)-m\xi_j(0)-(l-1)\bar{\xi}_{1,l-1}(0)+(l-1)\xi_l(0)\right]\\
&\overset{(*)}{\leq} \frac{1}{m}e^{-T}\left[\sum_{i=1}^{m}\xi_i(0)-\sum_{i=1}^{l-1}\xi_i(0)+(l-1-m)\xi_l(0)\right]\\
&= \frac{1}{m}e^{-T}\left[-\sum_{i=m+1}^{l-1}\xi_i(0)+(l-1-m)\xi_l(0)\right]\\
&\overset{(*)}{\leq} \frac{1}{m}e^{-T}\left[-(l-1-m)\xi_l(0)+(l-1-m)\xi_l(0)\right]\\
&= 0,
\end{aligned} \tag{3.42}$$

where inequalities $(*)$ derive from Hypothesis 2: since $j \leq l$, $\xi_j(0) \geq \xi_l(0)$. However, $(\bar{\xi}_{1,m}-\xi_j)(T) \leq 0$ contradicts that $\xi_j(T) < \xi_i(T)$ for every $i \in \{1,...,m\}$. From this we conclude that if $T \geq t_l$, then for every $i \in \{1,...,l\}$, $\xi_i(T) = \bar{\xi}_{1,l}(T)$ for an optimal control strategy fulfilling Hypothesis 3.

Let us now show that if $T < t_{l+1}$, then for every $k \in \{l+1,...,N\}$, $\alpha_i \equiv 0$ and $\xi_k(T) < \bar{\xi}_{1,l}(T)$.

$$\begin{aligned}
\bar{\xi}_{1,l}(T)-\xi_k(T) &\overset{(1)}{=} e^{-T}\left(\bar{\xi}_{1,l}(0)-\xi_k(0)-\int_0^T e^{\frac{N-1}{N}s}(\frac{1}{l}\sum_{j=1}^{l}\alpha_j-\alpha_k)(s)\bar{\xi}(0)ds\right)\\
&\overset{(2)}{\geq} e^{-T}\left(\bar{\xi}_{1,l}(0)-\xi_k(0)-\int_0^T e^{\frac{N-1}{N}s}\frac{1}{l}\bar{\xi}(0)ds\right)\\
&= e^{-T}\left(\bar{\xi}_{1,l}(0)-\xi_k(0)-\frac{N}{N-1}\frac{1}{l}\bar{\xi}(0)(e^{\frac{N-1}{N}T}-1)\right)\\
&\overset{(3)}{>} e^{-T}\left(\bar{\xi}_{1,l}(0)-\xi_k(0)-(\bar{\xi}_{1,l}(0)-\xi_{l+1}(0))\right)\\
&= e^{-T}\left(\xi_{l+1}(0)-\xi_k(0)\right)\\
&\overset{(4)}{\geq} 0,
\end{aligned} \tag{3.43}$$

where:

(1) was computed using the evolutions of ξ_k and $\bar{\xi}_{1,l}$: $\dot{\xi}_k = -\xi_k + (1-\alpha_k)\bar{\xi}$ and $\dot{\bar{\xi}}_{1,l} = -\bar{\xi}_{1,l} + (1-\frac{1}{l}\sum_{i=1}^{l}\alpha_i)\bar{\xi}$,

(2) was obtained from inequalities $\sum_{j=1}^{l}\alpha_j(t) \leq 1$ and $\alpha_k(t) \geq 0$ for all t,

(3) comes from the inequality: $T < t_{l+1} = \frac{N}{N-1}\ln(\frac{N-1}{N}l\frac{\bar{\xi}_{1,l}(0)-\xi_{l+1}(0)}{\bar{\xi}(0)}+1)$,

(4) derives from Hypothesis 2 since $k \geq l+1$.

Hence, for every $k \in \{l+1,...,N\}$, $\xi_k(T) \geq \bar{\xi}_{1,l}(T)$. Furthermore, the transversality condition (3.18) and Proposition 3.3.1 imply that for all $t \in [0,T]$ for every $i \in \{1,...,l\}$, $\lambda_k(t) < \lambda_i(t)$ and the

Pontryagin Maximum Principle as seen in Section 3.3 states that $\alpha_k \equiv 0$. So $\xi_k(T) \geq \bar{\xi}_{1,l}(T)$.

We proved that if $T \in [t_l, t_{l+1}[$, then for every $i \in \{l+1, ..., N\}$, $\alpha_i \equiv 0$. Since $\bar{\xi}$ is fully determined as $\alpha \in \mathcal{U}_{\text{FS}}$, this means that for all $i \in \{l+1, ..., N\}$, $\xi_i(T)$ is also fully determined (satisfying the equation $\dot{\xi}_i = -\xi_i + \bar{\xi}$). On the other hand, we proved that for all $i \in \{1, ..., l\}$, $\xi_i(T) = \bar{\xi}_{1,l}(T)$ and that $\sum_{i=1}^{l} \alpha_i \equiv 1$, so $\bar{\xi}_{1,l}$ is also fully determined (satisfying the equation $\dot{\bar{\xi}}_{1,l} = -\bar{\xi}_{1,l} + \frac{l-1}{l}\bar{\xi}$). Hence, any strategy such that for all $i \in \{1, ..., l\}$, $\xi_i(T) = \xi_{1,l}(T)$ with $\sum_{i=1}^{l} \alpha_i \equiv 1$ and for all $i \in \{l+1, ..., N\}$, $\alpha_i \equiv 0$ is optimal for Problem 2. □

Notice that this optimal control strategy is not sparse, as control is split among more and more agents as time goes. However, it is not unique and one could very well act on one agent at a time until all reach the known final velocities. Going back to the general Problem 1, we prove that under certain conditions, the optimal control strategy uses full strength at all time, i.e. $\alpha^{\text{opt}} \in \mathcal{U}_{FS}$.

Theorem 3.5.2 (Sufficient condition for full control)**.**
Define the time $t_N = \frac{N}{N-1} \ln\left(\frac{(N-1)^2}{N} \frac{\bar{\xi}_{1,N-1}(0) - \xi_N(0)}{\bar{\xi}(0)} + 1\right)$ *as in Theorem 3.5.1.*
If $T \geq t_N$*, then the optimal strategies* α^{opt} *to Problem 1 belong to* $\mathcal{U}_{FS}$ *and for these controls* $\xi_i(T) = \bar{\xi}(T)$ *for every* $i \in \{1, ..., N\}$*.*

Proof. If $T \geq t_N$, then the instantaneous decrease strategy designed in Theorem 3.5.1 is optimal. Indeed, we noticed that the migration functional can be written as the sum of two terms (3.11): $\mathbb{V} = \bar{\xi}^2 + \frac{1}{N}\sum(\xi_i - \bar{\xi})^2$. The strategy designed in Theorem 3.5.1 minimizes $\bar{\xi}(T)$ by using full control at all time, hence minimizing $\bar{\xi}(T)^2$ since $\bar{\xi} > 0$. Furthermore, it achieves $\xi_i(T) = \bar{\xi}(T)$ for all $i \in \{1, ..., N\}$, thus minimizing the second term $\frac{1}{N}\sum(\xi_i - \bar{\xi})^2$. Hence any optimal control strategy has to use full control at all time and achieve $\xi_i(T) = \bar{\xi}(T)$ for every $i \in \{1, ..., N\}$ in order to perform as well. □

We finally address the general case stated in Problem 1: minimize $\mathbb{V}(T)$ over the set of controls $\mathcal{U}$, for a given final time T. In the following theorem, we show the existence of an initial "Inactivation" time interval: the optimal strategy can require to let the system evolve freely (i.e. without control) at initial time, before acting on it with full strength.

Theorem 3.5.3 (Inactivation Principle)**.**
If $T < t_N$*, then one of the two holds: any control strategy* α^{opt} *either belongs to* $\mathcal{U}_{FS}$ *and the strategy designed in Theorem 3.5.1 is optimal, or there exists some* $\delta < T$ *such that* $\alpha^{opt} \equiv 0$ *on* $[0, \delta]$*, and* $\sum \alpha_i^{opt} \equiv 1$ *on* $[\delta, T]$*.*

Proof. According to Hypothesis 3, we can assume that $\xi_1(T) \geq \xi_i(T)$ for every $i \in \{1, ..., N\}$. Furthermore, $\bar{\xi}(T) > 0$, so $\xi_1(T) > 0$. From the transversality condition (3.18) we deduce: $\lambda_1(T) \geq \lambda_i(T)$ for every $i \in \{1, ..., N\}$ and $\lambda_1(T) > 0$. From Prop. 3.3.1, we know that for all $t \in [0, T]$, $\lambda_1(t) \geq \lambda_i(t)$. According to Prop. 3.3.2, full control is used at time t if $\lambda_1(t) > 0$ and no control is used if $\lambda_1(t) < 0$. Let us study the evolution of λ_1: $\dot{\lambda}_1 = \frac{1}{N}\sum \alpha_j \lambda_j - \bar{\lambda} + \lambda_1$. By the Pontryagin maximum principle, we always have $\sum \alpha_j \lambda_j \geq 0$. Furthermore, $\lambda_1 - \bar{\lambda} \geq 0$. So $\dot{\lambda}_1(t) \geq 0$ for all $t \in [0, T]$. We show that $\lambda_1 = 0$ at most at one point. Indeed, suppose that $\lambda_1(\tau) = 0$ for some $\tau \in [0, T]$ and that $\dot{\lambda}_1(\tau) = 0$. Then $\dot{\lambda}_1(\tau) = -\bar{\lambda}(\tau)$ so $\bar{\lambda}(\tau) = \lambda_1(\tau) = 0$, and since the λ_i's are ordered, $\lambda_i(\tau) = \bar{\lambda}(\tau)$ for every $i \in \{1, ..., N\}$. According to Proposition 3.3.1, $\lambda_i(t) = \bar{\lambda}(t)$ for all time t and every i. Since $\lambda_1(T) > 0$, there exists a time interval $[\tau^*, T]$ such that $\lambda_1(t) > 0$ for all $t \in [\tau^*, T]$. On this interval, $\dot{\lambda}_1 = \frac{1}{N}\lambda_1 \sum_j \alpha_j = \frac{1}{N}\lambda_1$, which gives: $\lambda_1(T) = \lambda_1(\tau^*)e^{\frac{1}{N}(T-\tau^*)}$. This contradicts the existence of a time τ at which $\lambda_1(\tau) = 0$. In conclusion, if $\lambda_1(\tau) = 0$, then $\dot{\lambda}_1(\tau) > 0$ so $\lambda_1 = 0$ at most at one point.

Hence, there is a dichotomy of cases:

Either $\lambda_1(t) \geq 0$ for all time, so $I(t) \neq \emptyset$ for all t, which implies that $\alpha^{\text{opt}} \in \mathcal{U}_{\text{FS}}$ according to Prop. 3.3.2. In this case, $\arg\max_{\alpha \in \mathcal{U}} \mathbb{V} = \arg\max_{\alpha \in \mathcal{U}_{\text{FS}}} \mathbb{V}$ and the control strategy designed in Theorem 3.5.1 for Problem 2 is optimal also for Problem 1.

Or there exists $\delta \in [0, T]$ such that $\lambda_1(t) < 0$ on $[0, \delta[$ and $\lambda_1(t) \geq 0$ on $[\delta, T]$, which implies that $\alpha(t) \equiv 0$ on $[0, \delta]$ and $\sum \alpha_i(t) \equiv 1$ on $]\delta, T]$. Practically, an optimal control strategy would consist in letting the system evolve without control on $[0, \delta[$. Then the full-control strategy from Theorem 3.5.1 can be applied on $[\delta, T]$ with the new initial positions $\xi(\delta)$. □

Remark 3.5.1. *Although this result may seem counter-intuitive, in certain cases it makes sense to let the system evolve freely, at least initially. Indeed, without control the system naturally regroups in order to reach consensus, minimizing $\sum_{i=1}^{N}(\xi_i - \bar{\xi})$ in (3.11), but keeping $\bar{\xi}$ constant. Actual examples of such cases are shown in the next section.*

Remark 3.5.2. *Note that a constraint $M < 1$ would not change the nature of the results. It would only mean acting with less strength on the controlled agents, therefore changing the values of the times t_l defined in Theorem 3.5.1, but the optimal control strategy would be unchanged. With a constraint $M > 1$, we can expect results similar to those of Section 3.4, with two kinds of Inactivation periods, consisting either in letting the system evolve freely, or in controlling it with a non-maximal total strength $0 < \sum_i \alpha_i < M$ (see Theorem 3.4.3 (ii) and (iii)).*

3.5.2 Practical Approach

We proved in the previous section that the optimal strategy can either be to act with full control as in Theorem 3.5.1, or to let the system evolve without control on some time interval $[0, \delta]$, before acting with full control on $]\delta, T]$. In this section, we explore the practicality of Inactivation strategies.

First, we run numerical simulations to find cases in which the optimal strategy involves Inactivation. The migration functional $\mathbb{V}_\delta$ can be computed explicitly as a function of δ. We then look for the value of δ that minimizes $\mathbb{V}_\delta(T)$. Let us denote by ξ^δ the solution to system (3.35) when no control is applied on $[0, \delta]$ and full control is used on $]\delta, T]$. Equation (3.35) gives:

$$\begin{cases} \dot{\xi}_i^\delta = -\xi_i^\delta + \bar{\xi}^\delta \\ \dot{\bar{\xi}}^\delta = 0 \end{cases} \qquad \text{on } [0, \delta], \tag{3.44}$$

which allows us to solve: $\xi_i^\delta(\delta) = e^{-\delta}\left(\xi_i^\delta(0) + \bar{\xi}^\delta(0)(e^\delta - 1)\right)$. We then apply the strategy designed in Theorem 3.5.1 with the new initial conditions $\xi^\delta(\delta)$ and the new final time $T - \delta$. Define the times $t_1^\delta = 0$ and for $l \in \{2, ..., N\}$, $t_l^\delta = \frac{N}{N-1} \ln\left((l-1)\frac{N-1}{N}\frac{\bar{\xi}_{1,l-1}^\delta(\delta) - \xi_l^\delta(\delta)}{\bar{\xi}^\delta(\delta)} + 1\right)$. Find $l \in \{1, ..., N-1\}$, such that $T - \delta \in [t_l^\delta, t_{l+1}^\delta[$. Then any strategy satisfying $\xi_i^\delta(T) = \bar{\xi}_{1,l}^\delta(T)$ for every $i \in \{1, ..., l\}$, $\sum_{i=1}^l \alpha_i(t) = 1$ for all $t \in [\delta, T]$, and $\alpha_i \equiv 0$ for every $i \in \{l+1, ..., N\}$ is optimal. From equation (3.35) we get:

$$\begin{cases} \dot{\bar{\xi}}_{1,l}^\delta = -\bar{\xi}_{1,l}^\delta + \frac{l-1}{l}\bar{\xi}^\delta \\ \dot{\xi}_i^\delta = -\xi_i^\delta + \bar{\xi}^\delta \quad \text{for } i \in \{l+1, ..., N\} \\ \dot{\bar{\xi}}^\delta = -\frac{1}{N}\bar{\xi}^\delta \end{cases} \qquad \text{on } [\delta, T], \tag{3.45}$$

from which we can solve:

$$\begin{cases} \xi_i^\delta(T) = \bar{\xi}_{1,l}^\delta(T) = e^{-(T-\delta)}\left(\bar{\xi}_{1,l}^\delta(\delta) + \frac{l-1}{l}\frac{N}{N-1}\bar{\xi}^\delta(0)(e^{\frac{N-1}{N}(T-\delta)} - 1)\right) & 1 \le i \le l, \\ \xi_i^\delta(T) = e^{-(T-\delta)}\left(\xi_i^\delta(\delta) + \frac{N}{N-1}\bar{\xi}^\delta(0)(e^{\frac{N-1}{N}(T-\delta)} - 1)\right) & l+1 \le i \le N. \end{cases} \tag{3.46}$$

We can now compute $\mathbb{V}^\delta(T) = \frac{1}{N}\sum_{i=1}^N \xi_i^\delta(T)^2$ and numerically look for $\min_{\delta \in [0,T]} \mathbb{V}^\delta(T)$ (see Figure 3.3).

Series of simulations were run to look for cases in which $\delta > 0$. Table 3.1 lists the percentage of such cases found over 1000 simulations, for different values of the number of agents and of the final time. Initial projected variables $\xi_i(0)$ were chosen randomly in the interval $[-1, 1]$ and such that the mean $\bar{\xi}$ is strictly positive. As expected (and proven in Theorem 3.5.2), for larger values of T, it is always optimal to act with full control at all time (in other words $\delta = 0$). One also notices that as

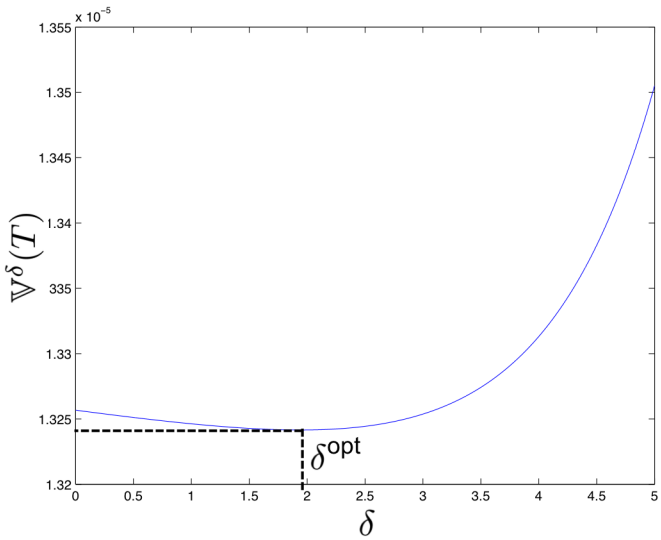


Figure 3.3: $V^{\delta}(T)$ with respect to Inactivation time δ. Here the optimal Inactivation time is $\delta = 1.94$.

the number of agents increases, “Inactivation" cases become less and less frequent.

Number of agents	5	10	20	50
T=3	1.6 %	0.9 %	0	0
T=4	1.8 %	0.7 %	0.3 %	0
T=5	1.0 %	0.2 %	0.2 %	0
T=6	0.2 %	0.1 %	0	0.1 %
T=7	0	0	0	0

Table 3.1: Percentage of cases in which $\delta > 0$ out of 1000 simulations. $\xi_i(0)$ chosen randomly in $[-1, 1]$.

Table 3.2 shows the average of the relative difference $\frac{\mathbb{V}_{\text{fc}} - \mathbb{V}^{\delta}}{\mathbb{V}_{\text{fc}}}$, where $\mathbb{V}^{\delta}$ was obtained by using optimal control and $\mathbb{V}_{\text{fc}}$ by using full control at all time (as designed in Theorem 3.5.1). The gain in performance when using the optimal strategy is minor (significantly less than 1% in most cases), and decreases as the number of agents increases.

Number of agents	5	10	20	50
T=3	0.073%	0.001%	-	-
T=4	0.27%	0.018%	0.001%	-
T=5	0.91%	0.056%	0.0069%	-
T=6	1.53%	0.2%	-	0.00003 %

Table 3.2: Average relative improvement of $\mathbb{V}^{\delta}$ w.r.t. $\mathbb{V}_{\text{fc}}$

The occurrence of Inactivation cases can be explained by looking at the two terms in the migration

functional $\mathbb{V} = \bar{\xi}^2 + \frac{1}{N}\sum(\xi_i - \bar{\xi})^2$ (3.11). When $\bar{\xi}^2$ is small, the control strategy should concentrate on minimizing the second term $\frac{1}{N}\sum(\xi_i - \bar{\xi})^2$, which does not necessarily require full control since the system naturally evolves to minimize this term. To confirm this reasoning, we look at the ratio $R := (\frac{1}{N}\sum(\xi_i - \bar{\xi})^2)/\bar{\xi}^2$ in one set of simulations ($N = 5$, $T = 3$) and find that the Inactivation cases correspond exactly to the largest values of R. Furthermore, the larger the ratio, the longer the Inactivation interval (see Figure 3.4).

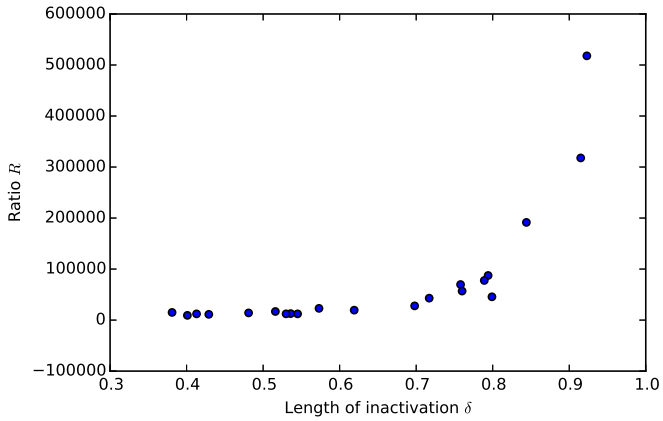


Figure 3.4: Ratio $R := (\frac{1}{N}\sum(\xi_i - \bar{\xi})^2)/\bar{\xi}^2$ as a function of the length of the Inactivation interval δ, for 20 simulations involving Inactivation with $N = 5$ and $T = 3$. The Inactivation δ increases as $\bar{\xi}^2$ tends to zero.

Hence, $\mathcal{U}_{\text{opt}} \cap \mathcal{U}_{\text{FS}} = \emptyset$ occurs in very few cases, namely those in which $\bar{\xi}^2 \ll \frac{1}{N}\sum(\xi_i - \bar{\xi})^2$. Furthermore, when Inactivation exists, the gain in performance compared to the full control strategy is very minor. For reasons of computational speed and complexity, it is very reasonable to neglect those cases and to apply the full control strategy at all time.

Figure 3.5 shows the evolution of the projected velocities $\xi_i,\ i \in \{1, ..., 10\}$ with respect to time, in a case where the optimal strategy requires full control at all time, with $T > t_{10}$. The control function is the one designed in Theorem 3.5.1 and acts first on ξ_1, then on ξ_1 and ξ_2, and so on until all have reached consensus (in terms of the projected velocities ξ_i), at which point it acts with equal strength on all agents to drive $\bar{\xi}$ down to 0.

3.6 Optimal control for integral cost

In this section we focus on minimizing the integral of the migration functional, with the constraint on the controls $M = 1$. As done in Section 3.5, we define two problems (where $\mathcal{U}$ (3.34) and $\mathcal{U}_{\text{FS}}$ (3.36) are defined as before).

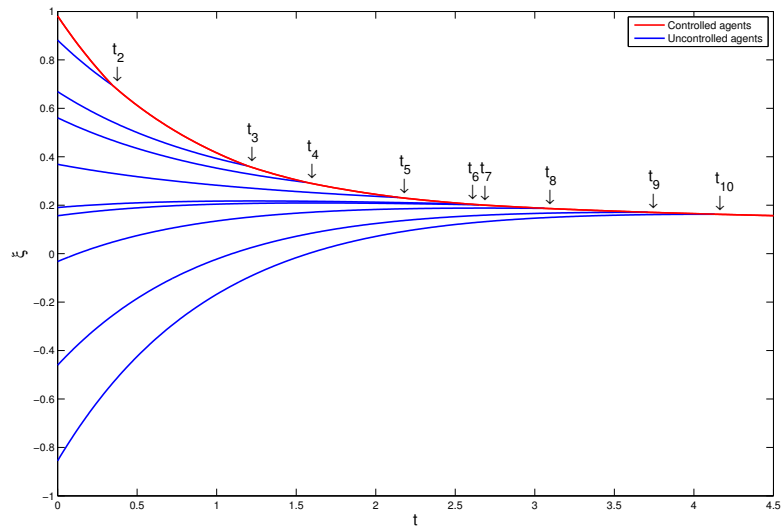


Figure 3.5: Evolution of the projected velocities ξ_i with the full strength optimal control for a system of 10 agents. In this example $\bar{\xi}(0) = 0.25$ so full control at all time is needed to drive $\bar{\xi}$ to the desired velocity $V = 0$ (i.e. $\delta = 0$). At final time $T = 4.5$ the system has reached consensus, but not yet at the desired velocity.

Problem 3. *Find* $\arg\min_{\alpha \in \mathcal{U}} \int_0^T \mathbb{V}(t)dt$.

Problem 4. *Find* $\arg\min_{\alpha \in \mathcal{U}_{FS}} \int_0^T \mathbb{V}(t)dt$.

3.6.1 Pontryagin's Maximum Principle

We first prove general results, with the aim of solving Problem 3. In order to use Pontryagin's maximum principle, we introduce the new Hamiltonian $H = \langle \lambda, f \rangle + \lambda^0 \mathbb{V}$ and the equations governing the covectors' evolution $\dot{\lambda}_i = -\frac{\partial H}{\partial \xi_i}$. Considering normal trajectories, we set $\lambda^0 = 1$ and obtain:

$$\begin{cases} H = \sum_{i=1}^{N} (-\lambda_i \xi_i) + \bar{\xi} \sum_{i=1}^{N} (1 - \alpha_i)\lambda_i + \sum_{i=1}^{N} \xi_i^2 \\ \dot{\lambda}_i = \lambda_i - \frac{1}{N} \sum_j (1 - \alpha_j)\lambda_j - 2\xi_i. \end{cases} \tag{3.47}$$

Since the final condition is not fixed, we have the following transversality condition for the covectors:

$$\lambda(T) = 0. \tag{3.48}$$

As in the minimization of the migration functional at final time (Section 3.3), we define I_λ and I_λ^+ (see (3.21)). Then minimizing $H = \sum_{i=1}^{N} -\alpha_i \lambda_i + \tilde{H}$ (where $\tilde{H}$ contains only uncontrolled terms)

requires the following : if $k \notin I_\lambda$, $\alpha_k = 0$; furthermore, if $I_\lambda^+ \neq \emptyset$, then $\sum_{i \in I_\lambda^+} \alpha_i = 1$.

As in Section 3.5, we make Hypothesis 2. Given the initial order on the agents' projected velocities ξ_i, we prove the following:

Lemma 3.6.1. *There exists an optimal control strategy satisfying:*

$$\forall t \in [0,T],\ \forall i,j \in \{1,...,N\},\ i<j \Rightarrow \xi_i(t) \geq \xi_j(t). \tag{3.49}$$

Proof. The proof is very similar to that of Lemma 3.3.1. Consider an optimal control strategy $\alpha \in \mathcal{U}$. Define $\tau = \sup\{t \mid \exists \beta \in \mathcal{U} \text{ s.t. } \int_0^T \mathbb{V}_\beta(s)ds = \int_0^T \mathbb{V}_\alpha(s)ds \text{ and } \xi^\beta \text{ satisfies (3.49) on } [0,t]\}$. Let us prove by contradiction that $\tau = T$. Suppose that $\tau < T$. Then there exist $i,j \in \{1,...,N\}$ with $i<j$ such that $\xi_i^\beta(\tau) = \xi_j^\beta(\tau)$ and $\xi_j^\beta(t) > \xi_i^\beta(t)$ on $]\tau, \tau+\delta]$ for some $\delta > 0$. Design a control strategy $\tilde{\beta}$ such that on $[\tau,T]$, $\tilde{\beta}_i = \beta_j$, $\tilde{\beta}_j = \beta_i$ and for every $k \in \{1,...,N\} \setminus \{i,j\}$, $\tilde{\beta}_k = \beta_k$. Then for all $t \in [\tau,T]$, $\xi_i^{\tilde{\beta}}(t) = \xi_j^\beta(t)$, and $\xi_j^{\tilde{\beta}}(t) = \xi_i^\beta(t)$. So for all $t \in [\tau, \tau+\delta]$, $\xi_i^{\tilde{\beta}}(t) \geq \xi_j^{\tilde{\beta}}(t)$ and for all $t \in [0,T]$, $\mathbb{V}^{\tilde{\beta}}(t) = \mathbb{V}^\beta(t)$. Proceeding likewise for every pair of indices (m,n) satisfying $m<n$ and $\xi_n^\beta(t) > \xi_n^\beta(t)$ on $]\tau,\tau+\delta]$ we are able to design a control strategy $\tilde{\beta}$ satisfying (3.49) on $[0,\tau+\delta]$ and $\int_0^T \mathbb{V}_\beta(t)dt = \int_0^T \mathbb{V}_\alpha(t)dt$, which contradicts the definition of τ. In conclusion, $\tau = T$, i.e. for all $t \in [0,T]$, for every $i,j \in \{1,...,N\}$, $i<j \Rightarrow \xi_i(t) \geq \xi_j(t)$. □

Hence, as in Section 3.5, we can assume Hypothesis 3: for all $t \in [0,T]$, if $i<j$, then $\xi_i(t) \geq \xi_j(t)$. By the following proposition, we shall prove that the same order is observed among the covectors λ_i.

Proposition 3.6.1.

$$\forall t \in [0,T],\ i<j \Rightarrow \lambda_i(t) \geq \lambda_j(t). \tag{3.50}$$

Proof. Let us reason by contradiction. Suppose that there exists $\tau \in [0,T]$ such that for some i<j, $(\lambda_i - \lambda_j)(\tau) < 0$. From the evolution of the covectors (3.47) we derive for all $t \geq \tau$: $(\lambda_i - \lambda_j)(t) = e^{t-\tau}\left((\lambda_i - \lambda_j)(\tau) - 2\int_\tau^t e^{-(s-\tau)}(\xi_i - \xi_j)(s)ds\right)$. Since $(\lambda_i - \lambda_j)(\tau) < 0$ and for all $s \in [0,T]$, $(\xi_i - \xi_j)(s) \geq 0$, we deduce that for all $t \in [\tau,T]$, $(\lambda_i - \lambda_j)(t) < 0$, which contradicts the final condition (3.48). □

Proposition 3.6.2. *Let $\tau \in [0,T]$ and $i,j \in \{1,...,N\}$, such that $(\lambda_i - \lambda_j)(\tau) = 0$. Then for all $t \geq \tau$, $(\lambda_i - \lambda_j)(t) = 0$ and $(\xi_i - \xi_j)(t) = 0$.*

Proof. Let $\tau \in [0,T]$ and $i,j \in \{1,...,N\}$, such that $(\lambda_i - \lambda_j)(\tau) = 0$. Then for all $t \geq \tau$,

$$(\lambda_i - \lambda_j)(t) = -2e^{t-\tau} \int_\tau^t e^{-(s-\tau)}(\xi_i - \xi_j)(s)ds. \tag{3.51}$$

Suppose for instance that $i < j$. According to Proposition 3.6.1, for all $t \in [0,T]$, $(\lambda_i - \lambda_j)(t) \geq 0$. Since we made Hypothesis 3, the right-hand side of equation (3.51) is nonpositive. This is only possible if both sides are equally zero. Hence, for all $t \geq \tau$, $(\lambda_i - \lambda_j)(t) = 0$ and $(\xi_i - \xi_j)(t) = 0$. □

The following proposition states that if at a certain point in time, two agents have the same projected velocities, then these should stay identical until final time.

Proposition 3.6.3. *Suppose that there exists $\tau \in [0,T]$ and $i,j \in \{1,...,N\}$ such that $\xi_i(\tau) = \xi_j(\tau)$. Then*

$$\text{for all } t \geq \tau,\ \xi_i(t) = \xi_j(t). \tag{3.52}$$

As a consequence, for almost all $t \geq \tau$, $\alpha_i(t) = \alpha_j(t)$.

Proof. Let $\tau \in [0,T]$ and $i,j \in \{1,...,N\}$. Define $\tilde{\tau} = \sup\{t \geq \tau \mid \xi_i(t) = \xi_j(t) \text{ for all } t \in [\tau, \tilde{\tau}]\}$. Notice from (3.35) that this implies that $\alpha_i(t) = \alpha_j(t)$ for almost every $t \in [\tau, \tilde{\tau}]$. Let us prove that $\tilde{\tau} = T$.

Suppose that $\tilde{\tau} < T$. Then there exists $\delta > 0$ such that for all $t \in]\tilde{\tau}, \tilde{\tau} + \delta]$, $\xi_i(t) \neq \xi_j(t)$. Define β such that $\beta = \alpha$ on $[0, \tilde{\tau}]$ and

$$\begin{cases} \beta_i = \beta_j = \frac{1}{2}(\alpha_i + \alpha_j) \\ \beta_k = \alpha_k \text{ for } k \neq i,\ k \neq j \end{cases} \qquad \text{on }]\tilde{\tau}, T], \tag{3.53}$$

and denote by ξ^β the corresponding trajectory. Notice that $\sum_k \alpha_k \equiv \sum_k \beta_k$, so according to (3.35), $\bar{\xi} \equiv \bar{\xi}^\beta$. This implies that $\xi_k = \xi_k^\beta$ for all $k \neq i,j$. Moreover, $\alpha_i + \alpha_j \equiv \beta_i + \beta_j$ so for all $t \in [\tilde{\tau}, T]$, $(\xi_i + \xi_j)(t) = (\xi_i^\beta + \xi_j^\beta)(t)$. Furthermore, ξ_i^β and ξ_j^β satisfy the same differential equation on $[\tilde{\tau}, T]$ and $\xi_i^\beta(\tau) = \xi_j^\beta(\tau)$, so for all $t \in [\tilde{\tau}, T]$, $\xi_i^\beta(t) = \xi_j^\beta(t) = \frac{1}{2}(\xi_i + \xi_j)(t)$. Define $\mathbb{V}_\alpha$ and $\mathbb{V}_\beta$ as the cost functions associated respectively with the controls α and β. Then $\mathbb{V}_\beta = \mathbb{V}_\alpha$ on $[0, \tilde{\tau}]$. On $]\tilde{\tau}, T]$,

$$\begin{aligned} \mathbb{V}_\alpha - \mathbb{V}_\beta &= \sum_k (\xi_k)^2 - \sum_k (\xi_k^\beta)^2 = (\xi_i)^2 + (\xi_j)^2 - (\xi_i^\beta)^2 - (\xi_j^\beta)^2 \\ &= (\xi_i)^2 + (\xi_j)^2 - 2(\frac{1}{2}(\xi_i + \xi_j))^2 = (\xi_i - \xi_j)^2. \end{aligned} \tag{3.54}$$

Hence, for all $t \in]\tilde{\tau}, \tilde{\tau} + \delta]$, $\mathbb{V}_\alpha(t) > \mathbb{V}_\beta(t)$, and for all $t \in [\tilde{\tau} + \delta, T]$, $\mathbb{V}_\alpha(t) \geq \mathbb{V}_\beta(t)$. We get

$\int_0^T \mathbb{V}_\beta < \int_0^T \mathbb{V}_\alpha$, which contradicts that α is an optimal control. In conclusion, $\tau = T$, which proves the proposition. □

3.6.2 Optimal full-strength control

We design an optimal control strategy for Problem 4:

Theorem 3.6.1. *Let $J(t) = \{i \in \{1, ..., N\} \mid \xi_i(t) = \max_j \xi_j(t)\}$. The following control α is optimal for Problem 4:*

$$\begin{cases} \forall i \in J(t),\ \alpha_i(t) = \frac{1}{|J(t)|} \\ \forall i \notin J(t),\ \alpha_i(t) = 0. \end{cases} \tag{3.55}$$

Proof. According to Pontryagin's maximum principle and the expression of the Hamiltonian (3.47), the optimal control strategy solving Problem 4 requires to set $\sum_{i \in I(t)} \alpha_i(t) = 1$ and $\alpha_k(t) = 0$ for $k \notin I(t)$, where $I(t) := \{i \mid \lambda_i(t) = \max_j \lambda_j(t)\}$. Furthermore, according to Proposition 3.6.2, if $\lambda_i(\bar{t}) = \lambda_j(\bar{t})$, then $\xi_i(t) = \xi_j(t)$ for all $t \geq \bar{t}$, and according to Proposition 3.6.3, $\alpha_i(t) = \alpha_j(t)$ for almost every $t \geq \bar{t}$. Hence, the optimal strategy in fact requires to set, for almost every $t \in [0, T]$,

$$\begin{cases} \forall i \in I(t),\ \alpha_i(t) = \frac{1}{|I(t)|} \\ \forall i \notin I(t),\ \alpha_i(t) = 0, \end{cases} \tag{3.56}$$

where $|\cdot|$ denotes the cardinality of a set. Let us prove that $I(t) = J(t)$ for almost every t. Assume that $i \in I(t)$ and (3.56) holds true. According to Proposition 3.6.1, the covectors are ordered, so $\lambda_1(t) = \cdots = \lambda_i(t)$. From Proposition 3.6.2 and Hypothesis 3, this implies $\xi_1(t) = \cdots = \xi_i(t)$, so $i \in J(t)$. Conversely, assume that $i \in J(t)$. Then from Hypothesis 3, $\xi_1(t) = \cdots = \xi_i(t)$. According to Proposition 3.6.3, $\alpha_1(t) = \cdots = \alpha_i(t)$. Since $\alpha(t)$ verifies (3.56), we deduce that $i \in I(t)$. Hence $I(t) = J(t)$ for almost every $t \in [0, T]$ and the optimal strategies (3.56) and (3.55) are equivalent. □

Notice that the control strategy in the case of integral cost minimization with full control (Problem 4) is equivalent to the Instantaneous decrease strategy of Prop. 3.2.1 (taking $M = 1$). It is more restrictive than the optimal strategy minimizing the final value of the migration functional with full control (Problem 2) seen in Section 3.5. Indeed, this control strategy cannot be sparse. In order to minimize $\int_0^T \mathbb{V}(t)dt$, one has to split the control among more and more agents. However, any optimal control solving Problem 4 is also optimal for Problem 2.

3.6.3 Optimal control in the general case

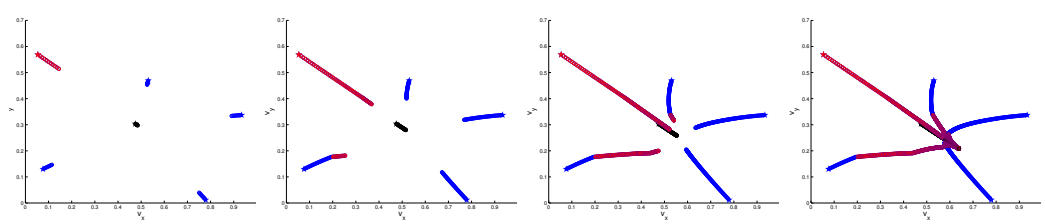
After designing the optimal strategy for Problem 4, we show that Problems 3 and 4 are actually equivalent, i.e. that the optimal control solving Problem 3 belongs to $\mathcal{U}_{\mathrm{FS}}$.

Theorem 3.6.2. *The optimal control strategy for Problem 3 requires using full-strength control, i.e.* $\alpha \in \mathcal{U}_{FS}$.

Proof. According to the Pontryagin Maximum Principle (see Section 3.6.1), if $\lambda_1(t) > 0$ for all t, then full control must be used at all time. Combining the final condition (3.48) and the evolution (3.47), we get $\lambda_1(T) = 0$ and $\dot{\lambda}_1(T) = -2\xi_1(T) < 0$. Hence there exists an interval $]t, T[$ on which $\lambda_1 > 0$. Let $\tau = \inf\{t \in [0, T] \text{ s.t. } \lambda_1(s) > 0 \text{ for all } s \in]t, T[\}$. Suppose that $\tau > 0$. Then $\lambda_1(\tau) = 0$. Furthermore, $\dot{\lambda}_1(\tau) = (\lambda_1 - \bar{\lambda})(\tau) - 2\xi_1(\tau)$. We compute: $\dot{\lambda}_1 - \dot{\bar{\lambda}} = \lambda_1 - \bar{\lambda} - 2(\xi_1 - \bar{\xi})$. Denoting $\Lambda = \lambda_1 - \bar{\lambda}$, we get the following evolution backwards in time: $\dot{\Lambda} = -\Lambda + 2(\xi_1 - \bar{\xi})$. Recall that backwards in time, we also have: $\dot{\xi}_1 = \xi_1 - (1-\alpha_1)\bar{\xi}$. If $\Lambda = \xi_1$, then $\dot{\Lambda} = \xi_1 - 2\bar{\xi} = \dot{\xi}_1 + (1-\alpha_1)\bar{\xi} - 2\bar{\xi} = \dot{\xi}_1 - (1+\alpha_1)\bar{\xi} < \dot{\xi}_1$. Since $\Lambda(T) = 0 < \xi_1(T)$, this implies that $\Lambda(t) < \xi_1(t)$ for all $t \in [\tau, T]$. Hence, $\dot{\lambda}_1(\tau) = \Lambda(\tau) - 2\xi_1(\tau) < 0$, which contradicts the definition of τ. We conclude that $\lambda_1(t) > 0$ for all $t \in]0, T[$, and that $\sum_i \alpha_i \equiv 1$. □

Hence, the control strategy designed in Theorem 3.6.1 is an optimal strategy for the minimization of integral cost (Prob. 3). Unlike in the minimization of the final cost (Prob. 1), there is no initial Inactivation period.

Figure 3.6 illustrates the control strategy designed in Theorem 3.6.1. In this example, 5 agents are to be controlled optimally to reach consensus at the target velocity $V = (1, 0)$. Initially (Figure 3.6a), only one agent is controlled, the agent with the biggest projected velocity over $\bar{v} - V$. The set $J(t) = \arg\max_{i \in \{1,\dots,N\}} \langle v_i, \frac{\bar{v}-V}{\|\bar{v}-V\|} \rangle$ contains more and more agents as time goes (3.6b, 3.6c) and eventually, control is split evenly among all agents (see Figure 3.6d).



(a) $t = 0.1$ (b) $t = 0.48$ (c) $t = 1.35$ (d) $t = 6.0$

Figure 3.6: Control of 5 agents to reach the target velocity $V = (1, 0)$. Agents are represented in the velocity space, controlled ones in red, uncontrolled ones in blue, and the mean velocity in black. Initial positions are marked by stars.

Chapter 4

Avoiding consensus: Black holes and declusterization

Introduction

The term "black swan" was first used by Nassim Nicholas Taleb in 2007 in his book *The Black Swan: The Impact of the Highly Improbable* [123], in which he focuses on the extreme impact of rare and unpredictable events. The "black swan theory" was since then developed to describe events that are extremely rare, have a massive impact, and are retrospectively predictable. One of the groundbreaking ideas of this recent theory is the fact that human behavior remains unpredictable. By focusing on what is known and probable, scientists tend to be surprised by major unexpected events. Taleb's philosophy requires one to accept the fact that there will always remain unknown factors - hence, one cannot make future predictions based only on the assumption of a population's rational behavior.

Bellomo et al. [7] have built upon this new theory, applying it to the context of social competition that can lead to extreme conflicts. Among an initially well distributed population, local social interactions can lead to unwanted clustering (of wealth, opinions, etc.). As pointed out in [130], individual behavior is often irrational: instead of making strategic decisions, individuals tend to imitate social neighbors. This behavior leads to clustering of opinions, or even consensus. Many models reproduce this phenomenon. In [130], this is modeled in a game-theoretic set-up, where agents play coordination games to improve their individual payoff. In the Voter model, agents imitate the action of a randomly selected counterpart [56]. In the Hegselmann-Krause (HK) bounded-confidence model, agents imitate others' behavior only if they are within a certain "confidence" radius [55]. In a competing approach, based on the so-called "topological" distance, agents imitate a given number of closest neighbors [6]. Another variation of the HK model consists of noticing that heterophilious dynamics enhance consensus [82]. Second-order models, such as the well-studied Cucker-Smale one

[31], may lead to alignment (i.e. agreement in the second variable) under suitable conditions on the interaction function [17, 49].

Self-organization has thus been extensively studied, especially focusing on the emergence of consensus or alignment that are an inherent property of certain dynamics. When consensus is not reached by the system, it is natural to ask whether it can be achieved by controlling the system, see for example [17, 18, 73]. Chapter 3 of this thesis also follows this line of work, by designing optimal control strategies to achieve consensus to a pre-determined state.

Here, we choose to study the opposite problem: given dynamics naturally leading to consensus under given conditions, we aim to control the system to avoid consensus, i.e. to keep the agents as far from one another as possible. In other words, we want to avoid a "black swan" phenomenon. Possible motivations include keeping a market from collapsing, or a crowd from converging to a localized dense conformation.

We study a first-order opinion model with a positive interaction function, and control the system via an additive feedback. We show that depending on the behavior of the interaction function $a(\cdot)$, several situations may arise. If $\lim_{s\to 0} sa(s) = +\infty$, there exists a "black hole" region, in which no control can keep the system from converging. On the other hand, if $\lim_{s\to 0} sa(s) = 0$, collapse to consensus can always be avoided. Far from the consensus manifold, we also observe two scenarios. If $\lim\limits_{s\to+\infty} sa(s) = 0$, there exists a "safety zone" in which the control can always keep the system far from consensus. This safety zone does not exist if $\lim\limits_{s\to+\infty} sa(s) = +\infty$.

We summarize these results in Table 4.1, giving criteria depending on $\alpha := \lim_{s\to\bar{s}} sa(s)$. The limit of $sa(s)$ when $s \to 0$ determines the existence of a black hole near the consensus manifold, that is a subset of $\mathbb{R}^{dN}$, containing the consensus manifold, that no control allows to escape from. On the other hand, the limit at infinity determines the existence of a safety zone far from the consensus manifold.

	$\bar{s} = 0$	$\bar{s} = \infty$
$\alpha = 0$	There exists a control strategy preventing consensus	There exists a **safety zone** far from consensus
$\alpha = \infty$	There exists a **black hole** (no strategy can avoid consensus for certain initial configurations)	There exists a **basin of attraction** (no safety zone far from consensus)

Table 4.1: Four different configurations determined by $\alpha = \lim_{s\to\bar{s}} sa(s)$

4.1 Preliminaries

Consider the general class of first-order control systems:

$$\dot{v}_i = f_i(v) + u_i, \qquad i = 1, \ldots, N, \quad v_i(t) \in \mathbb{R}^d, \tag{4.1}$$

where the dynamics f_i can be arbitrary.

The dynamics can be chosen to depend solely on the distance $v_i - v_j$, like in the well known Hegselman-Krause model of opinion dynamics:

$$\dot{v}_i = \frac{1}{N} \sum_{j=1}^{N} a(\|v_i - v_j\|)(v_j - v_i) + u_i. \tag{4.2}$$

Definition 4.1.1. *The state characterized by $v_1 = ... = v_N$ is referred to as **consensus**. We denote by $\mathcal{M}_{\mathrm{c}}$ the **consensus manifold** defined by:*

$$\mathcal{M}_{\mathrm{c}} := \{(v_i)_{i \in \{1,\ldots,N\}} \mid \forall (j,k) \in \{1, \ldots, N\}^2,\ v_j = v_k\}. \tag{4.3}$$

Remark 4.1.1. *The consensus state is an equilibrium for the Hegselman-Krause opinion dynamics without control (i.e. $u \equiv 0$). For this reason, system (4.2) is sometimes referred to as **consensus dynamics**.*

Without control and with a positive interaction potential $a(\cdot)$, the system converges to consensus. The aim of this work is to study under what conditions convergence to consensus can be avoided.

Notice that if at least two agents have different states, for instance if $v_i \neq v_j$ for some $i, j \in \{1, \ldots, N\}$, then the system is not in consensus.

Definition 4.1.2. *The system is said to **avoid consensus** if there exist $i, j \in \{1, \ldots, N\}$ such that $v_i \neq v_j$.*

However, avoiding consensus might still leave the system in a "dangerous" state, unwanted in some real-life situations. For instance, if each v_i represents an investor's decision, consensus might lead to a market crash. Whether or not the system is exactly in consensus state has little impact on the outcome: if one investor thinks differently than the mass (e.g. $v_j \neq v_1 = \ldots = v_{j-1} = v_{j+1} = \ldots = v_N$), it might not be enough to prevent a market collapse. With such applications in mind, we define a state of the system that is truly far from consensus, as the following:

Definition 4.1.3. *We say that the system is* ***dispersed****, or in* ***dispersion*** *state, if there exists* $\epsilon > 0$ *such that for all* $i, j \in \{1, \ldots, N\}$, $\|v_i - v_j\| \geq \epsilon$.

Notice that the condition "avoiding consensus" is weaker than the condition of dispersion. The system is said to avoid consensus if it is not in a neighborhood of the consensus manifold. More constraining, the condition of dispersion is satisfied if and only if the system is outside of a neighborhood of a larger manifold, that we refer to as the **clustering** manifold.

Definition 4.1.4. *The system is said to have* ***clusters*** *if there exist* $(i, j) \in \{1, \ldots, N\}^2$ *such that* $v_i = v_j$. *We denote by* $\mathcal{M}_{\mathrm{cl}}$ *the* ***clustering manifold****, defined by:*

$$\mathcal{M}_{\mathrm{cl}} := \{(v_i)_{i \in \{1,\ldots,N\}} \mid \exists (j,k) \in \{1, \ldots, N\}^2 \ s.t. \ v_j = v_k\}. \tag{4.4}$$

The consensus manifold $\mathcal{M}_{\mathrm{c}}$ is thus contained in the clustering manifold $\mathcal{M}_{\mathrm{cl}}$. More specifically, $\mathcal{M}_{\mathrm{c}}$ is a $d-$dimensional manifold embedded in $(\mathbb{R}^d)^N$, while $\mathcal{M}_{\mathrm{cl}}$ is a stratified set in the sense of Whitney (see Figure 4.1). We recall the definition of a stratified set:

Definition 4.1.5. *A set* $E \subset \mathbb{R}^n$ *is called* ***stratified*** *in the sense of Whitney if there exists a countable (locally finite) collection of pairwise disjoint manifolds* $(\mathcal{M}_i)_{i \in \mathbb{N}}$ *such that:*

1. $\mathcal{M}_i$ *is an embedded manifold of dimension* d_i
2. *If* $\mathcal{M}_i \cap \partial \mathcal{M}_j \neq \emptyset$, *then* $\mathcal{M}_i \subset \partial \mathcal{M}_j$ *and* $d_i < d_j$.

Moreover, we say that E *has separate strata if for every* $i \neq j$, *we have* $\mathcal{M}_i \cap \partial \mathcal{M}_j = \emptyset$.

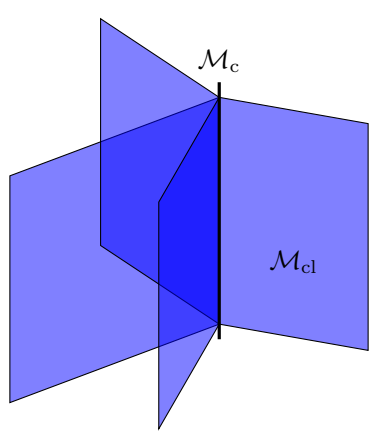


Figure 4.1: Schematic representation of the consensus manifold $\mathcal{M}_{\mathrm{c}}$ (black vertical line) contained in the stratified clustering manifold $\mathcal{M}_{\mathrm{cl}}$ (blue).

To characterize these different states, we introduce several functionals:

- the variance $V(t) = \frac{1}{2N^2} \sum_{i,j} \|v_i(t) - v_j(t)\|^2$
- the entropy functional $W(t) = \frac{1}{N^2} \sum_{i,j} \ln \|v_i(t) - v_j(t)\|$
- the generalized entropy functional $W_g(t) = \frac{1}{2N^2} \sum_{i,j} g(\|v_i(t) - v_j(t)\|^2)$ (with certain conditions on g).

Each of these functionals measures the distance of the system from either the consensus manifold or the clustering manifold. For example, the well-known variance characterizes the state of consensus:

Proposition 4.1.1. *Let $(v_i)_{i\in\{1,\dots,N\}} \in (\mathbb{R}^d)^N$, and let $V(t) = \frac{1}{2N^2} \sum_{i,j} \|v_i(t) - v_j(t)\|^2$. The system $(v_i(t))_{i\in\{1,\dots,N\}}$ is in the state of consensus if and only if $V(t) = 0$.*

4.1.1 A generalized entropy

Here we propose a general approach, which consists of designing feedback controls for the general class of first-order control systems (4.1).

To this aim, we define a generalized entropy functional W_g. In order to be able to characterize the dispersion of the system, we require specific properties for the function g.

Definition 4.1.6. *Let $g : \mathbb{R}^{+*} \to \mathbb{R}^+$ be a continuous, increasing function such that $\lim_{s\to 0} g(s) = -\infty$ and $\lim_{s\to+\infty} g(s) < \infty$. We define a generalized entropy functional W_g for system (4.9) as:*

$$W_g(t) = \frac{1}{2N^2} \sum_{i,j} g(\|v_i(t) - v_j(t)\|^2).$$

Remark 4.1.2. *The advantage of defining such an entropy functional is that we are able to characterize completely the dispersion of the system via the following condition:*

$$W_g > \eta \text{ if and only if } \exists \epsilon(\eta) \text{ s.t. } \forall i, j,\ \|v_i - v_j\| > \epsilon \tag{4.5}$$

Notice that the classical entropy functional W defined using $g(\cdot) := \ln(\cdot)$ does not allow such a characterization due to the fact that $\ln(s)$ grow without bounds when s tends to infinity.

Theorem 4.1.1. *Let $W_g = \frac{1}{2N^2} \sum_{i,j} g(\|v_i(t) - v_j(t)\|^2)$ be an entropy functional as defined in Definition 4.1.6. The following two statements are equivalent:*

1. *There exists $\eta > 0$ such that for all $t > 0$, $W_g(t) > \eta$*
2. *There exists $\varepsilon > 0$ such that for all $t > 0$, for all $i, j \in \{1, ..., N\}$, $\|v_i(t) - v_j(t)\| > \varepsilon$*

If the conditions above are satisfied, the system is dispersed at all time.

Proof. Let $W_g(t) > \eta$ for all $t > 0$. Suppose that for all $\varepsilon > 0$, there exist $t > 0$ and $i, j \in \{1, .., N\}$ such that $\|v_i(t) - v_j(t)\| \leq \varepsilon$. Let $C := \sup_{x>0} g(x)$. Let $A > (N^2 - 1)C - 2N^2\eta$. There exist $t > 0$ and $i, j \in \{1, .., N\}$ such that $g(\|v_i(t) - v_j(t)\|^2) < -A$. Then $W_g(t) \leq \frac{(N^2-1)C-A}{2N^2} < \eta$, which contradicts $W_g(t) > \eta$. The converse is trivial. □

4.1.2 Control strategy

We aim to design a feedback control strategy to keep the system in a dispersed state. From Theorem 4.1.1, maximizing W_g will ensure that the system is dispersed, hence that it is far from the state of consensus.

Given $M > 0$, we define the set of controls as

$$\mathcal{U}_M := \big\{u : [0, \infty) \to (\mathbb{R}^d)^N \;\big|\; u \text{ measurable}, \; \sum_{i=1}^{N} \|u_i\| \leq M\big\}. \tag{4.6}$$

The condition $\sum_{i=1}^{N} \|u_i\| \leq M$ is known as the $\ell_1^N - \ell_2^d$-norm constraint. It is known to promote the sparse behavior of the control (see [17]).

Let us start by computing $\dot{W}_g(t)$. Define $z_{ij} := v_i - v_j$ and note that $z_{ij} = -z_{ji}$. Since $\dot{z}_{ij} = f_i(v) - f_j(v) + u_i - u_j$, we get:

$$\begin{aligned}
\dot{W}_g =& \frac{1}{N^2} \sum_{1 \leq i < j \leq N} g'(\|z_{ij}\|^2) \langle z_{ij}, f_i(x) - f_j(x) + u_i - u_j \rangle \\
=& \frac{1}{N^2} \sum_{1 \leq i < j \leq N} g'(\|z_{ij}\|^2) \langle z_{ij}, f_i(x) + u_i \rangle - \frac{1}{N^2} \sum_{1 \leq i < j \leq N} g'(\|z_{ij}\|^2) \langle z_{ij}, f_j(x) + u_j \rangle \\
=& \frac{2}{N^2} \sum_{1 \leq i < j \leq N} g'(\|z_{ij}\|^2) \langle z_{ij}, f_i(x) + u_i \rangle \\
=& \frac{2}{N} \sum_{i=1}^{N} \langle \frac{1}{N} \sum_{j=1}^{N} g'(\|z_{ij}\|^2) z_{ij}, f_i(x) + u_i \rangle
\end{aligned} \tag{4.7}$$

Let $S_i := \frac{1}{N} \sum_{j=1}^{N} g'(\|z_{ij}\|^2) z_{ij}$. Let $i_0 := \arg\max_i \|S_i\|$, representing a weighted mean of influences of all agents on agent i. Then the control strategy maximizing $\dot{W}_g$ at all time t is sparse in the following sense:

$$u_i = \begin{cases} M \frac{S_i}{\|S_i\|} & \text{for } i = i_0 \\ 0 & \text{for all } i \neq i_0. \end{cases} \tag{4.8}$$

In particular this sparse control strategy applies to the Krause system (4.2) where

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} a(\|x_i - x_j\|)(x_j - x_i).$$

4.2 Main results

We now choose to focus our study on the first order consensus model with positive coefficients that are defined as functions of the state. Let $a \in \mathcal{C}^0(\mathbb{R}^+, \mathbb{R}^+)$ and $M > 0$. We define the controlled evolution of the system as follows:

$$\dot{v}_i = \frac{1}{N}\sum_{j=1}^{N} a(\|v_i - v_j\|)(v_j - v_i) + u_i, \tag{4.9}$$

where $u \in \mathcal{U}_M$ (see equation (4.6)).

4.2.1 The Black Hole

In Section 4.1.2, we designed a control strategy in the general case of system (4.1). We now study the more specific first-order consensus model (4.9). In this section, we prove that for certain potential functions $a(\cdot)$, there exists a "Black Hole zone", i.e. given a certain bound M on the control (with $\sum_{i=1}^{N} \|u_i\| \leq M$), for certain initial conditions, it is impossible to avoid convergence to consensus (the "Black Swan" phenomenon).

Theorem 4.2.1. *Let a be an attraction potential such that $\lim_{s\to 0} sa(s) = +\infty$. Then for all $M > 0$, there exists $\epsilon > 0$ such that if for all (i,j), $\|v_i(0) - v_j(0)\| < \epsilon$, then the system converges to consensus in finite time regardless of the control strategy.*

Proof. We study the evolution of the variance $V(t) = \frac{1}{2N^2}\sum_{i,j} \|v_i(t) - v_j(t)\|^2$.

$$\begin{aligned}
\dot{V} &= \frac{1}{2N^2}\sum_{i,j} 2\langle v_i - v_j, \dot{v}_i - \dot{v}_j\rangle \\
&= \frac{1}{N^2}\sum_{i,j}\langle v_i - v_j, \frac{1}{N}\sum_k a(\|v_i - v_k\|)(v_k - v_i) - \frac{1}{N}\sum_k a(\|v_j - v_k\|)(v_k - v_j) + u_i - u_j\rangle
\end{aligned}$$

The uncontrolled part of $\dot{V}$ writes:

$$
\begin{aligned}
&\frac{1}{N^2}\sum_{i,j}\langle v_i - v_j, \frac{1}{N}\sum_k a(\|v_i - v_k\|)(v_k - v_i) - \frac{1}{N^2}\sum_{i,j}\langle v_i - v_j, \frac{1}{N}\sum_k a(\|v_j - v_k\|)(v_k - v_j)\\
=&\frac{1}{N^3}\sum_{i,j,k}\left(\langle v_i - v_k, a(\|v_i - v_k\|)(v_k - v_i)\rangle + \langle v_k - v_j, a(\|v_i - v_k\|)(v_k - v_i)\rangle\right)\\
&-\frac{1}{N^3}\sum_{i,j,k}\left(\langle v_i - v_k, a(\|v_j - v_k\|)(v_k - v_j)\rangle + \langle v_k - v_j, a(\|v_j - v_k\|)(v_k - v_j)\rangle\right)\\
=&\frac{2}{N^3}\sum_{i,j,k}\langle v_i - v_k, a(\|v_i - v_k\|)(v_k - v_i)\rangle = \frac{2}{N^2}\sum_{i,k}\langle v_i - v_k, a(\|v_i - v_k\|)(v_k - v_i)\rangle\\
=&-\frac{2}{N^2}\sum_{i,k} a(\|v_i - v_k\|)\|v_k - v_i\|^2.
\end{aligned}
$$

Let $M > 0$. Since $\lim_{s\to 0} sa(s) = +\infty$, for all $A > 0$, there exists $\epsilon > 0$ such that for all $s < \epsilon$, $a(s) \geq \frac{A}{s}$. Near consensus, that is when for all i and j, $\|v_i(t) - v_j(t)\| \leq \epsilon$:

$$
\begin{aligned}
\dot{V} &= -\frac{2}{N^2}\sum_{i,j} a(\|v_i - v_j\|)\|v_i - v_j\|^2 + \frac{1}{N^2}\sum_{i,j}\langle v_i - v_j, u_i - u_j\rangle\\
&\leq -\frac{2}{N^2}A\sum_{i,j}\|v_i - v_j\| + \frac{1}{N^2}2M\sum_{i,j}\|v_i - v_j\|
\end{aligned}
$$

In particular, for $A = 2M$, in the corresponding region near consensus,

$$
\dot{V} \leq -\frac{2M}{N^2}\sum_{i,j}\|v_i - v_j\| \leq -\beta\sqrt{V}
$$

by equivalence of the norms with $\beta > 0$. Hence V tends to 0 in finite time. □

Remark 4.2.1. *The condition* $\lim_{s\to 0} sa(s) = +\infty$ *does not generalize to the integral condition on* a*:* $\int_0^{s_0} a(s)ds = +\infty$*. Take for instance* $a(s) = \frac{1}{s}$*. Then* $\int_0^{s_0} a(s)ds = +\infty$*, but* $\lim_{s\to 0} sa(s) = 1$*. Indeed, going back to the proof above, the derivative of the variance satisfies:*

$$
\dot{V} = \frac{1}{N^2}\sum_{i,j}\|v_i - v_j\| + \frac{1}{N^2}\sum_{i,j}\langle v_i - v_j, u_i - u_j\rangle \leq \frac{1 - 2M}{N^2}\sum_{i,j}\|v_i - v_j\|
$$

If $M < 1/2$*, then consensus is unavoidable, but for bigger values of* M *the possibility of acting on the system to prevent consensus remains.*

Theorem 4.2.1 shows that if the interaction between agents is very strong when they are close to each other (characterized by the condition $\lim_{s\to 0} sa(s) = +\infty$), then for every bound M on the control, there exists a zone close to the consensus manifold such that no control with bound M can

prevent consensus. We call this phenomenon the *Black Hole*. This is a local phenomenon. We now look at the behavior of the system far from the consensus manifold, that is when each pair of agents is sufficiently separated. We show in Sections 4.2.2 and 4.2.3 that depending on the strength of the decrease of a near infinity, there may or may not exist a *safety zone* far from the consensus manifold, that is a stable zone (given appropriate control).

4.2.2 Safety Zone

Here we give sufficient conditions on the potential for the existence of a *safety zone*. Given a bound M on the control, there exist initial conditions such that the control can always keep the system away far from consensus.

Theorem 4.2.2. *Let a be an attraction potential such that $\lim_{s\to+\infty} sa(s) = 0$. Construct W_g as in Definition 4.1.6. Then for all bound $M > 0$ on the control, there exists a safety zone in which $\dot{W}_g > 0$.*

Remark 4.2.2. *According to Theorem 4.1.1, the condition $\dot{W}_g > 0$ is enough to ensure that the system remains far from the consensus manifold at all time.*

Proof. From (4.7), we have:

$$\max_u \dot{W}_g = \frac{1}{N}\sum_{i=1}^{N}\langle S_i, \frac{1}{N}\sum_{k=1}^{N} a(\|v_i - v_k\|)(v_i - v_k)\rangle + \frac{M}{N}\|S_{i_0}\| \tag{4.10}$$

Since $\lim_{s\to+\infty} sa(s) = 0$, for all $\epsilon > 0$, there exists $\mu_1(\epsilon) > 0$ such that if for all i, j, $\|v_i - v_j\| \geq \mu_1(\epsilon)$, then $\frac{1}{N}\sum_{k=1}^{N} a(\|v_i - v_k\|)\|v_i - v_k\| \leq \epsilon$. Furthermore, due to the monotony of the chosen function g, for all $\bar{W}$, there exists $\mu_2(\bar{W}) > 0$ such that if $W_g \geq \bar{W}$, then for all i, j, $\|v_i - v_j\| \geq \mu_2(\bar{W})$.

Let $\epsilon < \frac{M}{N}$. Suppose that at $t = 0$ the agents are spread out enough that for all i, j, $\|v_i(0) - v_j(0)\| \geq \mu_2(W_g(0)) \geq \mu_1(\epsilon)$. Then $\max_u \dot{W}_g \geq \|S_{i_0}\|(\frac{M}{N} - \epsilon) \geq 0$. If we choose a control strategy maximizing $\dot{W}_g$ at all time, we ensure that for all $t > 0$, $W_g(t) \geq W_g(0)$, so that for all i, j, $\|v_i(t) - v_j(t)\| \geq \mu_2(W_g(0))$. □

This first theorem covers a wide range of interaction potentials. Given that the interaction potential $a(\cdot)$ decreases enough at infinity, we ensure the existence of a safety zone far from the consensus manifold. This for instance applies to potentials $a(\cdot)$ with compact support. However, notice that the interaction potential $a(s) = \frac{1}{s}$ does not meet the required conditions of Theorem 4.2.2. Here we state a new theorem dealing with functions that decrease at the speed of $1/s$.

Theorem 4.2.3. *Let g meet the required conditions for W_g to be a generalized entropy functional for system* (4.9)*, and let the interaction potential $a(\cdot)$ satisfy the following conditions*

(i) $g'(s)a(\sqrt{s})s \leq -g(s)$*;*

(ii) $g'(s)\sqrt{s}$ *and* $a(\sqrt{s})\sqrt{s}$ *have the same monotony.*

Then there exists a control for system (4.9) *such that if $W_g(0) \geq 0$, then $W_g(t) \geq 0$ for all $t \geq 0$. In other words, there exists a safety zone far from consensus.*

Proof. Let us study the time-evolution of the generalized entropy functional W_g. From (4.7), we have:

$$\dot{W}_g = \frac{2}{N}\sum_{i=1}^{N}\langle S_i, f_i\rangle + \frac{2}{N}\sum_{i=1}^{N}\langle S_i, u_i\rangle \tag{4.11}$$

where $S_i := \frac{1}{N}\sum_{k=1}^{N} g'(\|v_i - v_k\|^2)(v_i - v_k)$ and $f_i := \frac{1}{N}\sum_{k=1}^{N} a(\|v_i - v_k\|)(v_k - v_i)$. According to Chebychev's inequality, $(\frac{1}{N}\sum_{i=1}^{N} a_i)(\frac{1}{N}\sum_{i=1}^{N} b_i) \leq (\frac{1}{N}\sum_{i=1}^{N} a_i b_i)$ provided that the sequences (a_i) and (b_i) are ordered in the same way, i.e. for all $i < j$, $a_i \leq a_j$ and $b_i \leq b_j$. Assumption (ii) allows us to use Chebychev's inequality, so that we can write:

$$\begin{aligned}
|\langle S_i, f_i\rangle| &\leq \|S_i\|\,\|f_i\| \\
&\leq \frac{1}{N}\sum_{j=1}^{N} g'(\|v_i - v_j\|^2)\|v_i - v_j\|\ \frac{1}{N}\sum_{j=1}^{N} a(\|v_i - v_j\|)\|v_j - v_i\| \\
&\leq \frac{1}{N}\sum_{j=1}^{N} g'(\|v_i - v_j\|^2)a(\|v_i - v_j\|)\|v_i - v_j\|^2 \\
&\leq -\frac{1}{N}\sum_{j=1}^{N} g(\|v_i - v_j\|^2)
\end{aligned}$$

where we used assumption (i) for the last inequality. Summing over i, we get:

$$|\frac{1}{N}\sum_{i=1}^{N}\langle S_i, f_i\rangle| \leq -\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} g(\|v_i - v_j\|^2) = -2W_g.$$

This allows us to bound the first term of (4.7) from below:

$$\frac{2}{N}\sum_{i=1}^{N}\langle S_i, f_i\rangle \geq -2\,|\frac{1}{N}\sum_{i=1}^{N}\langle S_i, f_i\rangle| \geq 4W_g.$$

If we design a control strategy that satisfies $\sum_{i=1}^{N}\langle S_i, u_i\rangle \geq 0$, we can bound $\dot{W}_g$ from below: for all $t \geq 0$, $\dot{W}_g(t) \geq 4W_g(t)$. This implies that $W_g(t) \geq W_g(0)$ for all $t \geq 0$. □

Remark 4.2.3. *From the proof above, it is obvious that Conditions* (i) *and* (ii) *only need to be satisfied for s big enough.*

Remark 4.2.4. *The improvement of Theorem 4.2.3 over Theorem 4.2.2 lies in the limit case $a : s \mapsto \frac{1}{s}$. For instance, Theorem 4.2.3 can be applied to interaction potentials of the type:*

$$a(s) = \begin{cases} 1 \text{ for } 0 \leq s \leq 1 \\ \frac{1}{s} \text{ for } s > 1 \end{cases}$$

Indeed, taking $g : s \mapsto -\frac{1}{s}$, we have

(i) $g'(s)a(\sqrt{s})s \leq \frac{1}{s} = -g(s)$

(ii) for s big enough, $s \mapsto g'(s)\sqrt{s} = s^{-3/2}$ and $s \mapsto a(\sqrt{s})\sqrt{s} = s^{-1/2}$ are both decreasing. According to Theorem 4.2.3, there exists a safety zone far from the consensus region, even if a does not meet the hypotheses of Theorem 4.2.2.

Theorem 4.2.4. *Let a be an attraction potential such that*

- $-2 \leq (sa(s))' \leq 0$
- $\lim_{s \to 0} \int_s^{s_0} \frac{1}{a(\sqrt{\tau})\tau} d\tau = +\infty$ *for* $s_0 > 0$.

Then there exists an explicit safety zone.

Proof. Let us prove that we can find an entropy functional W_g such that a satisfies the conditions stated in Theorem 4.2.3. Let $s_0 > 0$ and a satisfying $-2 \leq (sa(s))' \leq 0$ and $\lim_{s \to 0} \int_s^{s_0} \frac{1}{a(\sqrt{\tau})\tau} d\tau = +\infty$ for $s_0 > 0$. Let g be defined by:

$$g(s) = -C \exp(-\int_{s_0}^{s} \frac{1}{a(\sqrt{\tau})\tau} d\tau)$$

for some positive constant C. This implies that

- g is increasing
- $\lim_{s \to 0} g(s) = \lim_{s \to 0} (-C \exp(\int_s^{s_0} \frac{1}{a(\sqrt{\tau})\tau} d\tau)) = -\infty$
- $\lim_{s \to 0} g(s) < \infty$ because $a(\cdot) > 0$

Thus W_g is a generalized entropy functional as defined in Definition 4.1.6. We now prove that $s \mapsto g'(s)\sqrt{s}$ and $s \mapsto a(\sqrt{s})\sqrt{s}$ are both decreasing.

$$(g'(s^2)s)' = -2s\frac{g'(s^2)}{a(s)s} + g(s^2)\frac{(a(s)s)'}{(a(s)s)^2} = \frac{g(s^2)}{(sa(s))^2}(2 + (sa(s))') \leq 0$$

Both conditions of Theorem 4.2.3 apply, hence there exists a security zone far from the consensus region. □

In sections 4.2.1 and 4.2.2, we showed the existence of a black hole zone near the consensus manifold if $\lim_{s\to 0} sa(s) = +\infty$ and the existence of a safety zone far from the consensus manifold if $\lim\limits_{s\to+\infty} sa(s) = 0$ (or $-2 \leq (sa(s))' \leq 0$ and $\lim\limits_{s\to 0} \int_s^{s_0} \frac{1}{a(\sqrt{\tau})\tau} d\tau = +\infty$). This suggests the existence of a "horizon" between safety and attraction to the black hole for interaction potentials that meet both conditions. The question remains of clarifying this horizon.

If the attraction potential does not satisfy the hypotheses of Theorems 4.2.2 or 4.2.3, we cannot ensure the existence of a safety zone. In fact, we show that in certain cases the safety zone does not exist and the whole space is a black hole, i.e. the black hole horizon is infinite.

Lemma 4.2.1. *If $a(s) = 1 + \frac{1}{s^2}$, there exists $M > 0$ such that the black hole horizon is infinite.*

Proof. Let $M \leq \frac{\alpha}{\sqrt{2}}$ for some $\alpha < 1$.
First assume that initially $\sqrt{V(0)} \leq \sqrt{2}M$ (i.e. some agents are already close to each other). We study the evolution of the variance $V = \frac{1}{2N^2}\sum_{i=1}^N \sum_{j=1}^N \|v_i - v_j\|^2$:

$$\frac{dV}{dt} = -\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N a(\|v_i - v_j\|)\|v_i - v_j\|^2 + \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \langle v_i - v_j, u_i - u_j\rangle. \tag{4.12}$$

The second term is related to V by equivalence of the norms:

$$\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \langle v_i - v_j, u_i - u_j\rangle \leq M\frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \|v_i - v_j\| \leq M\frac{1}{N^2}N\sqrt{\sum_{i=1}^N\sum_{j=1}^N \|v_i - v_j\|^2} = M\sqrt{2V}$$

Since $a(s) \geq \frac{1}{s^2}$, while $\sqrt{V} \leq \sqrt{2}M$ we have

$$\frac{dV}{dt} \leq -1 + M\sqrt{2V} \leq -1 + 2M^2 \leq \alpha - 1, \tag{4.13}$$

so V converges to 0 in finite time.
Let us now suppose that $\sqrt{V(0)} > \sqrt{2}M$, so the initial conformation is far from the consensus manifold. While this condition is satisfied, since $a(s) \geq 1$, we write:

$$\frac{dV}{dt} \leq -2V + M\sqrt{2V} = \sqrt{V}(-2\sqrt{V} + \sqrt{2}M) \leq -\sqrt{V}(\sqrt{2}M). \tag{4.14}$$

So V decreases until $\sqrt{V} = \sqrt{2}M$. When that happens, we are brought back to the first case. □

4.2.3 Basin of attraction

In Theorems 4.2.2 and 4.2.3, we saw that if $a(\cdot)$ decreases fast enough to 0 at infinity, then there exists a "safety" zone near infinity (i.e. when the agents are far from each other). Here we show that this safety zone does not always exist.

Theorem 4.2.5. *If $\lim_{s\to+\infty} sa(s) = +\infty$, then there is no "safety" zone far from consensus. In other words, there exist two sets B_1 and B_2 with $B_1 \subset B_2$, such that for all $v(0) \in \mathbb{R}^{dN} \setminus B_1$, for all control $u \in \mathcal{U}_M$, there exists $T > 0$ such that for all $t \geq T$, $v_u(t) \in B_2$.*

Proof. Let $A > M$. There exists $s_0 > 0$ such that if $s > s_0$, then $sa(s) > A$. Let $B_1 := \{(v_i)_{i\in\{1,...,N\}} \in \mathbb{R}^{dN} \,|\exists i,j \in \{1,...,N\}, \|v_i - v_j\| \leq s_0\}$. Then while $v \in B_1^c = \{(v_i)_{i\in\{1,...,N\}} \in \mathbb{R}^{dN} \,|\forall i,j \in \{1,...,N\}, \|v_i - v_j\| > s_0\}$, the variance V decreases as a quadratic function of time:

$$\dot{V} = -\frac{1}{N^2}\sum_{i,j} a(\|v_i - v_j\|)\|v_i - v_j\|^2 + \frac{1}{N^2}\sum_{i,j}\langle v_i - v_j, u_i - u_j\rangle \leq \frac{M-A}{N^2}\sum_{i,j}\|v_i - v_j\|$$

Since $A > M$, by equivalence of the norms, there exists $\gamma > 0$ such that $\dot{V} \leq -\gamma\sqrt{V}$. Hence V decreases until $v(T) \in B_1$ for some $T > 0$. When $v \in B_1$, it might become possible to act on the system again. If the control allows to obtain again $v \in B_1^c$, again V becomes strictly decreasing until $v \in B_1$. Hence for all $t > T$, $V(t) \leq \frac{s_0}{2}$. This implies that for all $t > T$, $\frac{1}{2N^2}\max_{i,j}\|v_i - v_j\| \leq \frac{s_0}{2}$. So for all $t > T$, $v(t) \in B_2$, where $B_2 := \{(v_i)_{i\in\{1,...,N\}} \in \mathbb{R}^{dN} \,|\forall i,j \in \{1,...,N\}, \|v_i - v_j\| \leq N^2 s_0\}$. $\square$

Remark 4.2.5. *In the case $d = 1$, $N = 2$, the consensus manifold is the line $v_1 = v_2$. The sets B_1 and B_2 are defined as: $B_1 = \{(v_1, v_2) \in \mathbb{R}^2 \mid \|v_1 - v_2\| \leq s_0\}$ and $B_2 = \{(v_1, v_2) \in \mathbb{R}^2 \mid \|v_1 - v_2\| \leq 4s_0\}$.*

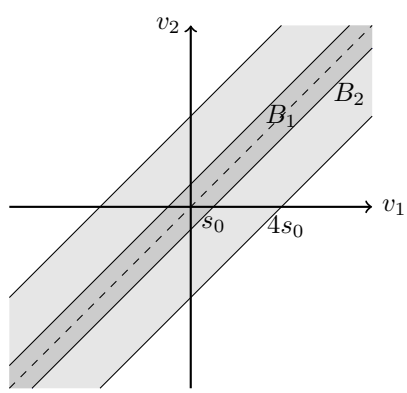


Figure 4.2: Consensus manifold (dashed line) and basin of attraction in the case $d = 1$, $N = 2$.

4.2.4 Collapse prevention

We saw in Section 4.2.1 that if $\lim_{s\to 0} sa(s) = +\infty$, there exists a Black Hole zone in which no control allows to avoid consensus.

On the other hand, we will show that if $\lim_{s\to 0} sa(s) = 0$, then consensus can always be avoided, in particular with the sparse control strategy defined in Section 4.1.2.

Theorem 4.2.6. *Suppose that* $\lim_{s\to 0} sa(s) = 0$. *Then the sparse control strategy defined in Section 4.1.2 prevents consensus. More specifically, there exists* $\bar{W}$ *and* $T > 0$ *such that for all* $t > T$, $W_g > \bar{W}$.

Proof. With the sparse control strategy, we have

$$\dot{W}_g = \frac{1}{N}\sum_{i=1}^{N}\langle S_i, \frac{1}{N}\sum_{k=1}^{N} a(\|v_i - v_k\|)(v_i - v_k)\rangle + \frac{M}{N}\|S_{i_0}\| \geq \|S_{i_0}\|(\frac{M}{N} - \max_{i,j} a(\|v_j - v_i\|)\|v_j - v_i\|).$$

Let $\varepsilon > 0$, with $\varepsilon < \frac{M}{N}$. Since $\lim_{s\to 0} sa(s) = 0$, there exists $\eta > 0$ such that for all $i, j \in \{1, ..., N\}$, $\|v_i - v_j\| \leq \eta \implies a(\|v_i - v_j\|)\|v_i - v_j\| \leq \varepsilon$. Suppose that the system is already close to consensus, so that for all $i, j \in \{1, ..., N\}$, $\|v_i - v_j\| \leq \eta$. Then $\dot{W}_g \geq (\frac{M}{N} - \varepsilon)\|S_{i_0}\| > 0$. W_g increases until there exists $i, j \in \{1, ..., N\}$, such that $\|v_i - v_j\| > \eta$. Denote by T the instant when that happens. While there exists $i, j \in \{1, ..., N\}$ such that $\|v_i - v_j\| > \eta$, $W_g \geq \frac{1}{2N^2}\max_{i,j} g(\|v_i - v_j\|^2) \geq \frac{1}{2N^2} g(\eta^2)$ due to the monotonicity of g. Hence for all $t > T$, $W_g(t) \geq \frac{1}{2N^2} g(\eta^2)$, which ensures boundedness away from consensus.

□

4.3 Numerical simulations

Numerical simulations illustrate the basin of attraction in the case $d = 1$, $N = 2$, see Figure 4.3. The interaction potential was chosen to be $a : s \mapsto s^{-1/2}$, so that $\lim_{s\to 0} sa(s) = 0$ and $\lim_{s\to\infty} sa(s) = +\infty$. Referring to the summarizing Table 4.1, we expect to exhibit the existence of a basin of attraction B_2, as well as to show that the control defined in Section 4.1.2 prevents convergence to consensus. We set the control bound $M = 1$. Choosing $A = M$ (the limit value in Theorem 4.2.5), we have $s_0 = 1$ and $B_1 := \{(v_1, v_2) \in \mathbb{R}^2 \mid |v_1 - v_2| \leq 1\}$ and $B_2 := \{(v_1, v_2) \in \mathbb{R}^2 \mid |v_1 - v_2| \leq 4\}$. Figure 4.4 shows the evolution of the entropy functional W_g for a choice of function $g : s \mapsto 1 - \frac{1}{s}$.

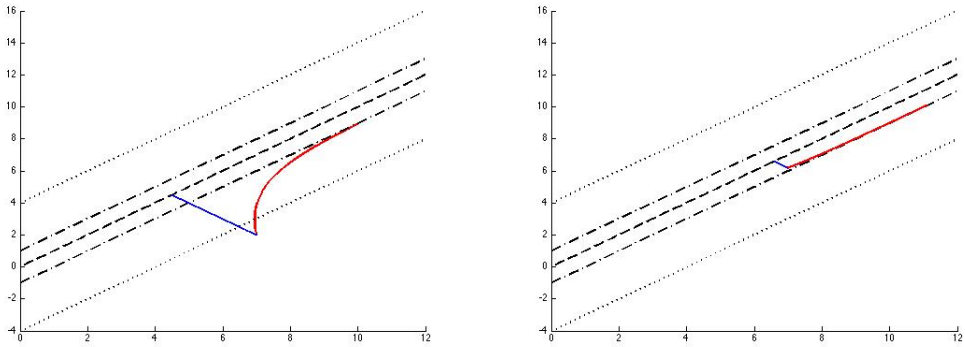


Figure 4.3: Evolution of $(v_1, v_2) \in \mathbb{R}^2$ in the case $a : s \mapsto s^{-1/2}$ with control (red) and without control (blue). Left: Initial configuration $(v_1, v_2) \notin B_2$. Right: Initial configuration $(v_1, v_2) \in B_1$. In both cases, without control the system tends to consensus (i.e.(v_1, v_2) tends to the consensus manifold $M = \{(v_1, v_2) \in \mathbb{R}^2 \mid v_1 = v_2\}$. With control, when the initial configuration is close to the consensus manifold M, the control is able to bring it away to a safer zone (right). When $(v_1(0), v_2(0))$ is initially far from consensus, despite the control, the system converges to a basin of attraction (left).

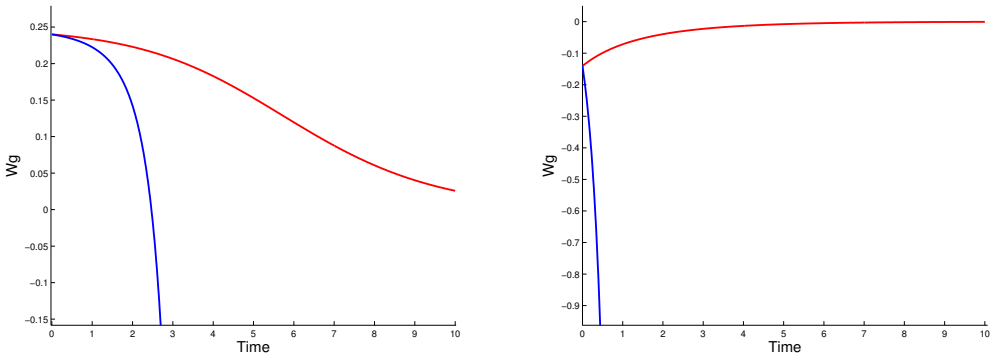


Figure 4.4: Evolution of $W_g(t)$ in the case $a : s \mapsto s^{-1/2}$ with control (red) and without control (blue). Left: Initial configuration $(v_1, v_2) \notin B_2$. Right: Initial configuration $(v_1, v_2) \in B_1$.

Chapter 5

Social dynamics models on general Riemannian manifolds

Introduction

The emergence of a group's global behavior from local interactions among individual agents is a fascinating feature of opinion dynamics. When local rules imply global patterns in a population, we are observing a phenomenon called *self-organization*. Traditionally, interest focuses on understanding the complex rules of interacting opinions which lead to certain global configurations, such as classic *consensus*, *alignment*, *clustering*, or the less studied *dancing equilibrium* [19]. For instance, in bounded-confidence models such as the one proposed by Hegselmann and Krause, the radius of interaction determines the clustering of the system [55]. Motsch and Tadmor studied the influence of the shape of the interaction potential on the convergence to consensus of the Hegselmann-Krause system [82]. Ha, Ha and Kim looked at the Cucker-Smale second-order alignment model and provided a condition on the interaction potential ensuring convergence of the system to alignment [49]. Cristiani, Frasca and Piccoli studied the effect of anisotropic interactions on the behavior of the group [28].

The dynamics of an opinion formation system greatly depend on the state-space [3]. Models on the Euclidean space in one dimension (for opinion dynamics) or in two or three dimensions (with applications to groups of animals or robots) have been extensively studied and are well understood. However, such models are locally linear, which may be a limitation when one strives to capture more complex phenomena and better represent reality [116]. In this line of thought, the Kuramoto model on the sphere $\mathbb{S}^1$ addresses the problem of synchronizing a large number of oscillators [68, 119]. There exist numerous applications to this model [36, 108, 111, 112]. Similarly, applications to satellite or ground vehicle coordination have motivated the development of models on special orthogonal groups [106, 107]: satellites evolve on $SO(3)$ while ground vehicles evolve on $SE(2)$ or $SE(3)$. A nonlinear

model of opinion formation on the sphere was also developed in [19].

The present work defines a general model of opinion dynamics on a Riemannian manifold. We investigate how the manifold on which the model is defined affects the global configurations resulting from opinion dynamics. These are the first steps to build a robust theory of opinion dynamics on general Riemannian manifolds.

There is an inherent difficulty in defining opinion dynamics on a general Riemannian manifold. Using the Riemannian distance, an agent will move towards a point by following the manifold's geodesics, which are well defined only locally. On a larger scale, there might not exist a unique geodesic. Another challenge is the extreme complexity of computing geodesics, even on a relatively simple manifold such as the torus [46]. One way around this issue is to consider the embedding of the manifold into a Euclidean space. Each agent's velocity is defined by projection of the other agents' influence onto the tangent space at that point. This is the choice made in [19].

Other than the mentioned practical aspect, there is an intrinsic rationale for choosing one approach over the other. When evolving along the geodesics of the manifold, one assumes that each agent has a global understanding of the manifold's geometry and is able to choose the shortest path among all possible ones. On the other hand, the approach based on the projection of the desired destination onto the tangent space implies that each agent only holds local information about the space in which it evolves. It chooses to move in the direction which locally seems to bring it closer to the target.

We explore these two specific approaches for our generalized model. The first method, "Approach A", uses projections in the Euclidean space in which the manifold is embedded. The second method, "Approach B", uses only geodesics defined on the manifold to define strength and direction of interaction. We exhibit properties of the interaction matrix that lead to specific kinds of equilibria. Simulations and examples compare the two methods. "Dancing equilibria" for approach B is shown (dancing equilibria was studied for approach A in [19].

We use the sphere and torus as example manifolds to evaluate these approaches. Specifically, we simulate dynamics on the following manifolds: $\mathbb{S}^1, \mathbb{S}^2$ and $\mathbb{T}^2$. These examples allow us to directly compare the two approaches, and see if one is more appropriate for a given manifold. We show the influence of the manifold's geometry on the dynamics by examining the dynamics resulting from the same interaction matrix in $\mathbb{S}^2$ and $\mathbb{T}^2$ and $\mathbb{R}^2$.

Opinion dynamics trajectories can resemble $n-$body choreography, that is, solutions to the well known $n-$body problem. These dynamics drive agents along orbits which may be shared by multiple agents. We refer to opinion dynamics trajectories along such orbits as "Social Choreography"

and investigate initial conditions and properties of the interaction matrix which give rise to Social Choreography, specifically in $\mathbb{R}^2$.

5.1 Choice of model

This work will primarily discuss two approaches to define opinion dynamics on a Riemannian manifold. Let M be a Riemannian manifold. Let $N \in \mathbb{N}$ represent the number of agents with opinions evolving on M. We denote by $x := (x_i)_{i\in\{1,...,N\}} \in M^N$ the set of opinions. For each $i \in \{1, ..., N\}$, $\dot{x}_i \in T_{x_i}M$. The opinions x_i evolve according to the following general dynamics:

$$\dot{x}_i = \sum_{j=1}^{N} a_{ij}\Psi(d(x_i, x_j))\nu_{ij} \tag{5.1}$$

where

- $a_{ij} \in \mathbb{R}$ is the interaction coefficient of the pair of agents i and j,
- $\Psi : \mathbb{R} \to \mathbb{R}$ is the interaction potential,
- $d(\cdot, \cdot) : M \times M \to \mathbb{R}^+$ represents the difference between opinions,
- $\nu_{ij} \in T_{x_i}M$ is a unit vector giving the direction of the influence of j over i.

Each of these terms is further specified in the following.

5.1.1 Approaches

The evolution of each agent's opinion depends on the opinions of all other agents, with influences weighted by the interaction coefficients a_{ij}. More specifically, an agent x_j's influence on x_i is determined by two elements: the direction of influence $\nu_{ij} \in T_{x_i}M$ and the magnitude of influence $\Psi(d(x_i, x_j)) \in \mathbb{R}^+$. We propose and study two different approaches for the choices of d and ν_{ij}. Approach A uses the embedding of M in $\mathbb{R}^n$ to define $d(x_i, x_j)$, whereas Approach B is intrinsic to M, with distance and direction of influence based on geodesics.

Approach A. Assume that M of dimension m is embedded in a Euclidean space $\mathbb{R}^n$, with $n \geq m$. Agent x_j acts on agent x_i via a projection onto $T_{x_i}M \subset \mathbb{R}^m$. Now considering points $(x_i, x_j) \in M^2$ as points of $\mathbb{R}^n$, the difference $x_j - x_i$ is a vector of $\mathbb{R}^n$. Given a vector subspace Y of $\mathbb{R}^n$, we denote

by $\Pi_Y y$ the projection of $y \in \mathbb{R}^n$ onto $Y \subset \mathbb{R}^n$ and define $d_P(\cdot,\cdot)$ as follows:

$$d_P(x_i, x_j) = \|\Pi_{T_{x_i}M}(x_j - x_i)\| \tag{5.2}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$. The same projection also defines the direction of influence of x_j on x_i:

$$\nu_{ij} = \begin{cases} \frac{\Pi_{T_{x_i}M}(x_j - x_i)}{\|\Pi_{T_{x_i}M}(x_j - x_i)\|} & \text{if } \Pi_{T_{x_i}M}(x_j - x_i) \neq 0 \\ 0 \quad \text{otherwise.} \end{cases} \tag{5.3}$$

With the specific choice $\Psi \equiv \mathrm{Id}$, system (5.1)-(5.2)-(5.3) becomes:

$$\dot{x}_i = \sum_{j=1}^{N} a_{ij} \Pi_{T_{x_i}M}(x_j - x_i). \tag{5.4}$$

This is the approach used in [19], applied to the sphere $\mathbb{S}^2$.

Notice that the magnitude of influence, $d_P(x_i, x_j)$, is symmetric for the sphere in the sense that $d_P(x_i, x_j) = d_P(x_j, x_i)$, but not symmetric for a general Riemannian manifold (see Figure 5.1). However, it is a continuous function defined for all pairs of points $(x_i, x_j) \in M^2$. The originality of this approach is that the influence of x_j on x_i is not related to a notion of distance between the points. The use of the projection of $x_j - x_i$ onto $T_{x_i}M$ reflects the concept of "local visibility." For the situation of two agents evolving on a one dimensional manifold, if $x_j - x_i \perp T_{x_i}M$, then a local displacement of x_i does not affect the distance between the points $\|x_i - x_j\|$. Indeed, a first order Taylor expansion gives: $x_i(\varepsilon) = x_i(0) + \varepsilon \dot{x}_i(0) + o(\varepsilon)$.

Supposing that x_j is fixed, we have:

$$\|x_i(\varepsilon) - x_j\|^2 = \langle x_i(\varepsilon) - x_j, x_i(\varepsilon) - x_j \rangle = \langle x_i(0) - x_j, x_i(0) - x_j \rangle + 2\varepsilon \langle \dot{x}_i(0), x_i(0) - x_j \rangle + o(\varepsilon)$$

so if $x_j - x_i(0) \perp T_{x_i(0)}M$, then $\|x_i(\varepsilon) - x_j\|^2 = \|x_i(0) - x_j\|^2 + o(\varepsilon)$. Hence if x_i only has local visibility, all directions of displacement seem equivalent (at first order), which justifies the influence of x_j over x_i to be zero if their difference is orthogonal to the tangent space of M at x_i. This is illustrated in Figure 5.1.

Approach B. This second approach defines d and ν_{ij} using the manifold M itself, and does not require any reference to the space in which M is immersed. This would make approach B a natural

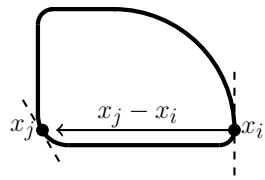


Figure 5.1: An example of a manifold M such that $d_p(x_i, x_j) \neq d_p(x_j, x_i)$, Using system (5.4), an agent is subject to "local visibility", and movement of x_i along $T_{x_i}M$ (dashed line through x_i) will not bring x_i closer to x_j in this local sense.

way to define system dynamics, however the complete knowledge of the geodesics between any two points on the manifold may be unrealistic. Furthermore, the geometry of the manifold may introduce difficulties to the uniqueness of ν_{ij}, particularly at the cut-locus of a point.

Definition 5.1.1. *The **cut locus** of a point $q \in M$ is the set of points $\mathcal{CL}(q) \subset M$ for which there are multiple geodesics between q and $p \in \mathcal{CL}(q)$ (see also [21]).*

Let $\gamma_{ij} : [0, 1] \to M$ denote a geodesic connecting x_i to $x_j, \gamma_{ij}(0) = x_i$ and $\gamma_{ij}(1) = x_j$. We then define the distance between x_j and x_i as the length of a geodesic, i.e. denoting by $g_y : T_yM \times T_yM \to \mathbb{R}^+$ the Riemannian metric at point $y \in M$,

$$d_G(x_i, x_j) = \int_0^1 \sqrt{g_{\gamma_{ij}(s)}(\dot{\gamma}_{ij}(s), \dot{\gamma}_{ij}(s))}ds. \tag{5.5}$$

The direction of influence is determined by the same geodesic:

$$\nu_{ij} = \begin{cases} 0 & \text{if } x_j = x_i \text{ or if } x_j \in \mathcal{CL}(x_i) \\ \frac{\dot{\gamma}_{ij}(0)}{\sqrt{g_{x_i}(\dot{\gamma}_{ij}(0), \dot{\gamma}_{ij}(0))}} & \text{otherwise.} \end{cases} \tag{5.6}$$

Unlike in Approach A, the magnitude of influence is a symmetric function: $d_G(x_i, x_j) = d_G(x_j, x_i)$. Furthermore, this approach ensures that the magnitude of influence of one agent on another is a function of the exact Riemannian distance between the agents.

Interaction networks. In finite-dimensional systems such as system (5.1), the set of interacting agents can be described by vertices of a graph. A directed edge exists from a vertex i to a vertex j if and only if $a_{ij} \neq 0$. The system depends on the interaction network, and likewise, if the coefficients a_{ij} are chosen to be functions of the state, the interaction network may change as a result of the dynamics. Two main types of interaction networks have been proposed in the literature: metric interactions and topological interactions. If interactions between agents occur only locally, only the

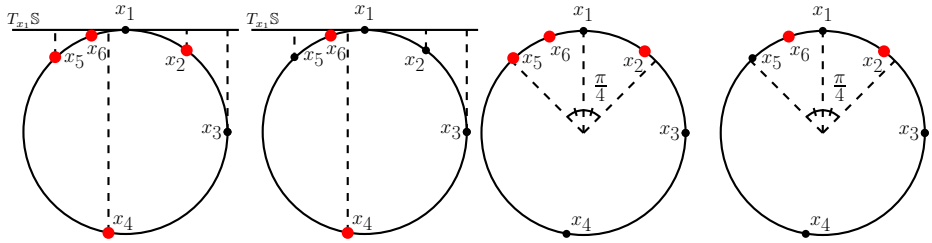
neighbors of agent i influence agent i. Metric interactions define the set of neighbors of agent i, given a radius $r > 0$, as

$$\mathcal{S}_i^r(x) = \{j \in \{1, \ldots, N\}, d(x_i, x_j) \leq r\}, \tag{5.7}$$

where $d(\cdot, \cdot)$ can represent either the projection or the geodesic distance, as specified in each of the two approaches described above (see equations (5.2) and (5.5)). The other main type of interactions specifies that an agent is influenced by only its k closest neighbors. We call these topological interactions [3]. We define the relative separation between two agents as $\alpha_{ij} = \mathbf{card}\{k : d(x_i, x_k) \leq d(x_i, x_j)\}$, The set of neighbors of agent i is then defined as the set of its k closest neighbors, i.e. for a given $k \in \mathbb{N}$,

$$\mathcal{S}_i^k(x) = \{j \in \{1, \ldots, N\}, \alpha_{ij} \leq k\}. \tag{5.8}$$

Figures 5.2 and 5.3 illustrate differences between the metric and topological networks for the specific example of $\mathbb{S}^1$, with each of the approaches A and B.



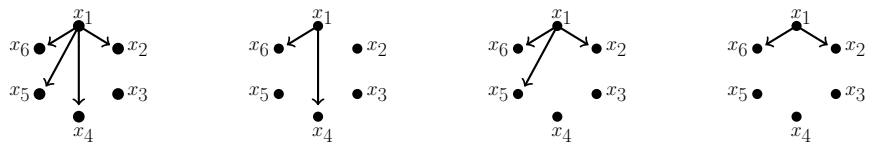
(a) Approach A metric, $r = \frac{\pi}{4}$

(b) Approach A topological, $k = 2$

(c) Approach B metric, $r = \frac{\pi}{4}$

(d) Approach B topological, $k = 2$

Figure 5.2: The set of agents that influence x_1 depends on how the interaction network is defined. In (a) and (b) the dashed lines show the projection of agents onto the tangent space of x_i, ($T_{x_i}\mathbb{S}^1$). The agents depicted in red with larger dots influence x_1. With the same configuration on $\mathbb{S}^1$, four combinations are possible (approach {A,B} type {Metric, Topological}). Each combination implies x_1 interacts with a different set of agents.



(a) Approach A metric, $r = \frac{\pi}{4}$ (b) Approach A topological, $k = 2$ (c) Approach B metric, $r = \frac{\pi}{4}$ (d) Approach B topological, $k = 2$

Figure 5.3: The agent x_1 is influenced by different agents depending on how the interaction network is defined. These networks may change as the dynamics move the agents on $\mathbb{S}^1$. Each agent $x_j, j \in \{1, \ldots, 6\}$ will have a network describing which other agents influence x_j. The interaction networks corresponding to systems from Figure 5.2.

Resolution of discontinuities. The definitions of ν_{ij} for approaches A and B (given by equations (5.3) and (5.6)) allow discontinuities of ν_{ij} at certain points. Thus, one must impose conditions on the interaction potential $\Psi \in \mathcal{C}^0(\mathbb{R}^+, \mathbb{R}^+)$, in order to ensure the continuity of the right-hand side of the system (5.1), and hence the existence and uniqueness of a solution. Table 5.1 lists the discontinuities of ν_{ij} and gives necessary conditions on Ψ to ensure the continuity of $\Psi(d(x_i, x_j))\nu_{ij}$.

Firstly, notice that in both approaches, ν_{ij} is discontinuous at the point $x_i = x_j$. Indeed, if $x_i = x_j$, $\nu_{ij} = 0$, whereas almost everywhere else, $\|\nu_{ij}\| = 1$. To ensure the continuity of $\Psi(d(x_i, x_j))\nu_{ij}$ at this point, we impose the following condition:

$$\Psi(0) = 0. \tag{5.9}$$

In approach A, we created a discontinuity of ν_{ij} at the points $x_j \in \mathcal{N}(x_i)$, where we denote by $\mathcal{N}(q)$ the set $\mathcal{N}(q) := \{q \in M \mid \Pi_{T_pM}(q - p) = 0\}$. For convenience of notation, we will use interchangeably the notations $\mathcal{N}(x_i)$ and $\mathcal{N}_i$. More specifically, we have $\lim_{x_j \to \mathcal{N}_i} \|\nu_{ij}\| = 1$ but $\|\nu_{ij}\| = 0$ if $x_j \in \mathcal{N}_i$ (see also Table 5.1). However, from the definition of d_P (see equation (5.2)), we have $\lim_{x_j \to \mathcal{N}_i} d_P(x_i, x_j) = 0$ and $d(x_i, x_j) = 0$ for $x_j \in \mathcal{N}_i$. Hence a sufficient condition for $\Psi(d(x_i, x_j))\nu_{ij}$ to be continuous is again:

$$\Psi(0) = 0. \tag{5.10}$$

In approach B, there is a discontinuity for $x_j \in \mathcal{CL}(x_i)$. Denoting by $B^{\text{geo}}(p, \rho)$ the geodesic ball of center p and radius ρ, we require the following condition on the influence function Ψ:

$$\Psi(d) = 0 \text{ for all } d \geq \epsilon \tag{5.11}$$

where $\epsilon := \inf\{\rho > 0 \mid \forall p \in M, B^{\text{geo}}(p,\rho) \cap \mathcal{CL}(p) = \emptyset\}$. This distance ϵ, also known as injectivity radius, is known to exist and be greater than 0 for any compact Riemannian manifold (see [21]).

Approach	A	B	A and B
Critical points	$x_j \in \mathcal{N}_i$	$x_j \in \mathcal{CL}(x_i)$	$x_j = x_i$
Discontinuities	$\lim_{x_j \to \mathcal{N}_i} \|\nu_{ij}\| = 1$ $\|\nu_{ij}\| = 0$ for $x_j \in \mathcal{N}_i$	$\lim_{x_j \to \mathcal{CL}(x_i)} \|\nu_{ij}\| = 1$ $\|\nu_{ij}\| = 0$ for $x_j \in \mathcal{CL}(x_i)$	$\lim_{x_j \to x_i} \|\nu_{ij}\| = 1$ $\|\nu_{ii}\| = 0$
Condition on Ψ	$\Psi(0) = 0$	$\Psi(d) = 0$ for all $d \geq \epsilon$	$\Psi(0) = 0$

Table 5.1: Possible discontinuities of the right-hand side of (5.1). The bottom row of the table show conditions for Ψ so that the system is continuous.

Notice that in the case of the geodesics approach (B), the condition $\Psi(d) = 0$ for all $d \geq \epsilon$ is incompatible with the use of the topological network (5.8). Indeed, if agent j is among the k closest neighbors of agent i, the topological network would require: $a_{ij} \neq 0$. However, the interaction between i and j would be canceled if $d_G(x_i, x_j) > \epsilon$. On the other hand, the metric interaction network as defined by (5.7) is compatible with approach A, and with approach B if the interaction radius is smaller than the injectivity radius: $r \leq \epsilon$. For simplicity purposes, in the rest of this chapter, we will consider that the interaction coefficients a_{ij} are constant, thus not requiring the need to differentiate between metric and topological networks.

5.1.2 Definitions and general results

Definition 5.1.2. *The configuration $x_1 = ... = x_N$ is called* ***consensus****. On the sphere, $\mathbb{S}^n$, A configuration such that, for every $j \in \{2, \ldots, N\}$, either $x_j = x_1$ or $x_j = -x_1$, which is not a concensus is called* ***antipodal equilibrium****.*

Proposition 5.1.1. *The consensus configuration is an equilibrium for system* (5.1)*.*

Proof. In both approaches A and B, if $x_i = x_j$, then $\nu_{ij} = 0$. Hence if $x_1 = ... = x_N$, then for all $i \in \{1, \ldots, N\}$, $\dot{x}_i = 0$. □

Proposition 5.1.2. *Let $N > d + 1$. Then for every $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_N) \in M^N$, there exists a square matrix $A = (a_{ij})_{i,j \in \{1,\ldots,N\}}$ such that $\bar{x}$ is an equilibrium for system* (5.1)*.*

Proof. The configuration $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_N)$ is an equilibrium if and only if

$$\frac{d}{dt}\bar{x}_i = \sum_{j=1}^{N} a_{ij}\Psi(d(\bar{x}_i, \bar{x}_j))\nu_{ij} = 0.$$

This is a system of at most Nd equations in the $N^2 - N$ unknowns a_{ij}, $i \neq j$, notice that $\Psi(d(x_i, x_i)\nu_{ii} = 0$, and diagonal values of A do not change the system. So if $N > d + 1$ there exists a nontrivial choice of the interaction coefficients for which $\bar{x}$ is an equilibrium. □

Definition 5.1.3. *The kinetic energy of System* (5.1)-(5.2)-(5.3) *is the quantity*

$$E(t) := \frac{1}{2} \sum_{i=1}^{N} \|\dot{x}_i(t)\|^2. \tag{5.12}$$

The kinetic energy of System (5.1)-(5.5)-(5.6) *is the quantity*

$$E(t) := \frac{1}{2} \sum_{i=1}^{N} g_{x_i}(\dot{x}_i(t), \dot{x}_i(t)). \tag{5.13}$$

Proposition 5.1.3. *Let M be a general Riemannian bounded manifold. Consider the dynamics given by projection onto the tangent space (Approach A) given by* (5.4)*. If the interaction matrix $A = (a_{ij})_{i,j \in \{1,\ldots,N\}^2}$ is symmetric, then*

$$\lim_{t \to \infty} E(t) = 0. \tag{5.14}$$

Proof. Let $F(t) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} \|x_i - x_j\|^2$. Using the symmetry of A, we prove that

$$\frac{d}{dt} F(t) = 4E(t). \tag{5.15}$$

Indeed, notice that

$$\nabla_{x_i}(\sum_{j=1}^{N} a_{ij} \|x_i - x_j\|^2) = 2\Pi_{T_{x_i}M} \sum_{j=1}^{N} a_{ij}(x_j - x_i) = 2\dot{x}_i.$$

Then we compute

$$
\begin{aligned}
\frac{d}{dt}F(t) &= \sum_{k=1}^{N}\langle \nabla_{x_k}\frac{1}{2}\sum_{i,j=1}^{N} a_{ij}(\|x_i-x_j\|^2), \dot{x}_k\rangle \\
&= \sum_{k=1}^{N}\langle \nabla_{x_k}\left[\frac{1}{2}\sum_{i=1}^{N} a_{ik}(\|x_i-x_k\|^2)+\frac{1}{2}\sum_{j=1}^{N} a_{kj}(\|x_k-x_j\|^2)\right], \dot{x}_k\rangle \\
&= \sum_{k=1}^{N}\langle 2\nabla_{x_k}\frac{1}{2}\sum_{i=1}^{N} a_{ik}(\|x_i-x_k\|^2), \dot{x}_k\rangle = \sum_{k=1}^{N}\langle 2\Pi_{T_{x_k}M}\sum_{j=1}^{N} a_{kj}(x_j-x_k), \dot{x}_k\rangle \\
&= 2\sum_{k=1}^{N}\|\dot{x}_k\|^2 = 4E(t).
\end{aligned}
\tag{5.16}
$$

where the third equality uses the property: $a_{ij} = a_{ji}$ for all i, j.

Since $E(t) \geq 0$, $F(t)$ is a non-decreasing function. Moreover $F(t)$ and $\frac{d^2}{dt}F(t)$ are bounded, since M is a bounded manifold. Hence $\frac{d}{dt}F(t) \to 0$ when $t \to \infty$, which implies that $\lim_{t\to\infty} E(t) = 0$. □

Remark 5.1.1. *Propositions 5.1.2 and 5.1.3 are generalizations of results proven for the case $M = \mathbb{S}^2$ in [19].*

Definition 5.1.4. *Let x solve the differential equation* (5.1)*. A* dancing equilibrium *is a configuration in which for all pairs of agents (i, j), the distance $\|x_i - x_j\|$ (in approach A) or $d_G(x_i, x_j)$ (in approach B) is constant.*

Remark 5.1.2. *This definition is a generalization of the concept of* dancing equilibrium *described in [19].*

Remark 5.1.3. *It follows immediately from definition 5.1.4 that the kinetic energy of a system in dancing equilibrium is constant.*

5.2 Analysis and simulations on $\mathbb{S}^1$

5.2.1 Models

We study both approaches A and B in the case $M = \mathbb{S}^1$, i.e. for the one-dimensional sphere embedded in $\mathbb{R}^2$. Let $(\theta_i)_{i\in\{1,\dots,N\}} \in [0, 2\pi]^N$ such that for all $i \in \{1, ..., N\}$, $x_i = (\cos\theta_i, \sin\theta_i)^T$.

Approach A. The projection onto an agent's tangent space can be rewritten as:

$$\begin{aligned}\Pi_{T_{x_i}} \sum_{j=1}^{N} a_{ij}(x_j - x_i) &= \sum_{j=1}^{N} a_{ij} \langle \begin{pmatrix} \cos\theta_j \\ \sin\theta_j \end{pmatrix} - \begin{pmatrix} \cos\theta_i \\ \sin\theta_i \end{pmatrix}, \begin{pmatrix} -\sin\theta_i \\ \cos\theta_i \end{pmatrix} \rangle \begin{pmatrix} -\sin\theta_i \\ \cos\theta_i \end{pmatrix} \\ &= \sum_{j=1}^{N} a_{ij}(-\sin\theta_i \cos\theta_j + \sin\theta_j \cos\theta_i) \begin{pmatrix} -\sin\theta_i \\ \cos\theta_i \end{pmatrix} \\ &= \sum_{j=1}^{N} a_{ij} \sin(\theta_j - \theta_i) \begin{pmatrix} -\sin\theta_i \\ \cos\theta_i \end{pmatrix}.\end{aligned} \tag{5.17}$$

So System (5.1)-(5.2)-(5.3) becomes:

$$\text{for all } i \in \{1, \dots, N\}, \quad \dot{\theta}_i \begin{pmatrix} -\sin\theta_i \\ \cos\theta_i \end{pmatrix} = \sum_{j=1}^{N} a_{ij} \Psi(|\sin(\theta_j - \theta_i)|) \operatorname{sgn}(\sin(\theta_j - \theta_i)) \begin{pmatrix} -\sin\theta_i \\ \cos\theta_i \end{pmatrix} \tag{5.18}$$

where $\operatorname{sgn}(\cdot)$ is the sign function defined by:

$$\text{for all } x \in \mathbb{R}, \quad \operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0. \end{cases} \tag{5.19}$$

We can then specify:

$$\text{for all } (i,j) \in \{1, \dots, N\}^2, \quad d_P(x_i, x_j) = |\sin(\theta_j - \theta_i)|, \quad \nu_{ij}^P = \operatorname{sgn}(\sin(\theta_j - \theta_i)). \tag{5.20}$$

This gives the system of scalar equations:

$$\text{for all } i \in \{1, \dots, N\}, \quad \dot{\theta}_i = \sum_{j=1}^{N} a_{ij} \Psi(|\sin(\theta_j - \theta_i)|) \operatorname{sgn}(\sin(\theta_j - \theta_i)). \tag{5.21}$$

In particular, in the case $\Psi \equiv \mathrm{Id}$, the system becomes the Kuramoto model [68].

$$\text{for all } i \in \{1, \dots, N\}, \quad \dot{\theta}_i = \sum_{j=1}^{N} a_{ij} \sin(\theta_j - \theta_i). \tag{5.22}$$

Approach B. For $M = \mathbb{S}^1$, the geodesics distance d_G and the vector ν_{ij}^G are given by:

$$d_G(x_i, x_j) = \arccos(\cos(\theta_j - \theta_i)) \quad , \quad \nu_{ij}^G = \operatorname{sgn}(\sin(\theta_j - \theta_i)). \tag{5.23}$$

System (5.1)-(5.5)-(5.6) is written:

$$\text{for all } i \in \{1, \dots, N\}, \quad \dot{\theta}_i = \sum_{j=1}^{N} a_{ij} \Psi(\arccos(\cos(\theta_j - \theta_i))) \operatorname{sgn}(\sin(\theta_j - \theta_i)). \tag{5.24}$$

In order for the system to be well defined, the interaction function Ψ must satisfy the conditions given in Table 5.1. Notice that the injectivity radius is constant over $\mathbb{S}^1$, with $\epsilon = \pi$. Possible choices involve choosing Ψ from a family of function defined as follows:

$$\Psi^a(d) = \begin{cases} \frac{1}{a} d & \text{for } d \leq a \\ \frac{d-\pi}{a-\pi} & \text{for } d > a \end{cases} \tag{5.25}$$

where $a \in (0, \pi)$.

Another possible choice is: $\Psi : x \mapsto \sin(x)$. Notice that for the specific choices $\Psi = \mathrm{Id}$ for approach A and $\Psi : x \mapsto \sin(x)$ for approach B, the two approaches A and B are equivalent.

5.2.2 Analysis

We first examine the different equilibria for both approaches.

Theorem 5.2.1. *Consider approach A, System* (5.21)*. Let $N \in \mathbb{N}$ be even. Suppose that for all $i \in \{1, \dots, N\}$ for all $j \in \{1, \dots \frac{N}{2}\}$, $a_{ij} = a_{i(j+\frac{N}{2})}$. Then any configuration that is centrally symmetric, i.e.*

$$\textit{for all } j \in \{1, \dots, \frac{N}{2}\},\ \theta_{j+\frac{N}{2}} = \theta_j + \pi$$

is an equilibrium.

Proof. Using the hypotheses from Theorem 5.2.1, we can easily compute:

$$\begin{aligned}
\dot{\theta}_i &= \sum_{j=1}^{N} a_{ij}\Psi(\|\sin(\theta_j-\theta_i)\|)\,\mathrm{sgn}(\sin(\theta_j-\theta_i)) \\
&= \sum_{j=1}^{N/2}[a_{ij}\Psi(\|\sin(\theta_j-\theta_i)\|)\,\mathrm{sgn}(\sin(\theta_j-\theta_i)) + a_{i(j+\frac{N}{2})}\Psi(\|\sin(\theta_{j+\frac{N}{2}}-\theta_i)\|)\,\mathrm{sgn}(\sin(\theta_{j+\frac{N}{2}}-\theta_i))] \\
&= \sum_{j=1}^{N/2}[a_{ij}\Psi(\|\sin(\theta_j-\theta_i)\|)\,\mathrm{sgn}(\sin(\theta_j-\theta_i)) + a_{ij}\Psi(\|\sin(\theta_j+\pi-\theta_i)\|)\,\mathrm{sgn}(\sin(\theta_j+\pi-\theta_i))] \\
&= 0.
\end{aligned}$$

□

Interestingly, Theorem 5.2.1 is not applicable to approach B. We illustrate the different behaviors of the two systems by studying the specific example of four agents initially in a rectangular configuration. According to Theorem 5.2.1, this configuration is an equilibrium for approach A, independently of the choice of interaction function Ψ. However, one can easily prove that in the geodesics-based approach B, with $N = 4$ and the choice $\Psi := \Psi^a$ with $a = \frac{3\pi}{4}$, the only equilibrium for which all agents have pairwise distinct positions is obtained by a regular polygon, i.e. all agents are evenly spaced out on the circle. This is illustrated by numerical simulations shown in Figure 5.4.

This highlights the fundamentally different behaviors of the systems (5.1)-(5.2)-(5.3) and (5.1)-(5.5)-(5.6) in the case $M = \mathbb{S}^1$.

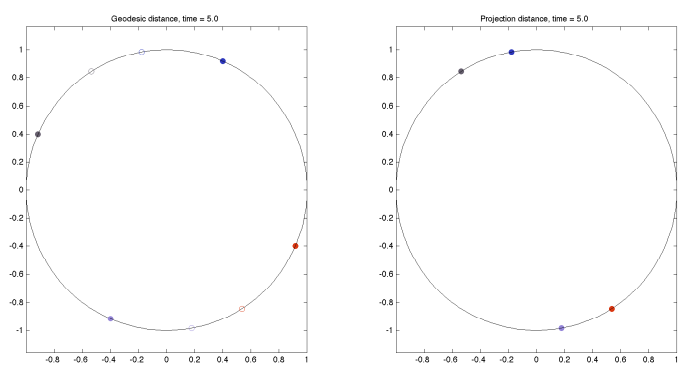


Figure 5.4: Initial (empty circles) and final positions (filled circles) of 4 agents initially on the vertices of a rectangle with approach B (left) and approach A (right), with $A = \nVdash$, $\Psi \equiv \mathrm{Id}$ (approach A) and $\Psi = \Psi^{3\pi/4}$(approach B) (see equation (5.25)). Notice that with approach A, initial and final positions are identical since any rectangle configuration is an equilibrium. However, with approach B, the system reaches a square configuration, the only possible equilibrium with pairwise distinct positions.

In both approaches A and B, conditions on the interaction matrix A can be found such that the

system forms a dancing equilibrium (see Definiton 5.1.4).

Theorem 5.2.2. *Consider the dynamics on* $\mathbb{S}^1$ *given by:*

$$\text{for all } i \in \{1, \dots, N\}, \quad \dot{\theta}_i = \sum_{j=1}^{N} a_{ij} \Psi(d(x_i, x_j)) \nu_{ij} \tag{5.26}$$

where $d(\cdot, \cdot)$ *and* ν *are given either by Approach A* (5.20) *or Approach B* (5.23)*. Let* $C \in \mathbb{R}$ *and suppose that for all* $i \in \{1, \dots, N\}$*,*

$$a_{ij} = \begin{cases} \frac{C}{\Psi(d(x_i(0), x_j(0)))} \nu_{ij} & \text{if } \Psi(d(x_i(0), x_j(0))) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.27}$$

Then the system is in a dancing equilibrium.

Proof. If the interaction matrix satisfies (5.27), then at $t = 0$,

$$\text{for all } i \in \{1, \dots, N\}, \quad \dot{\theta}_i(0) = \sum_{j=1}^{N} C = CN$$

so for all $(i, j) \in \{1, \dots, N\}^2$, $\dot{\theta}_i(0) - \dot{\theta}_j(0) = 0$. Then $d(x_i, x_j)$ does not change in time, and (5.27) holds for all time. □

Numerical simulations show the evolution of the system (5.26) with condition (5.27) for the projection or the geodesic distance, see Figures 5.5 and 5.6.

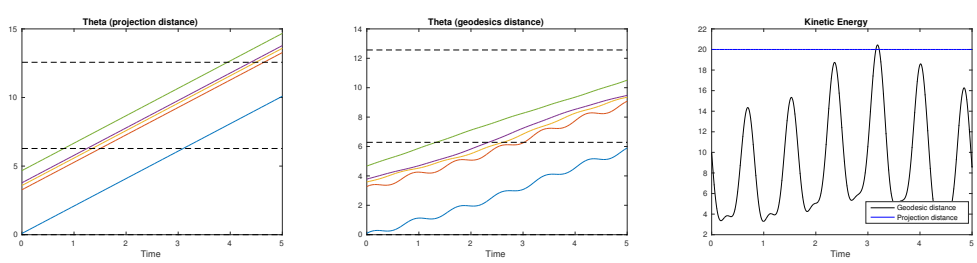


Figure 5.5: Evolution of the system (5.26) with Approach A (left) Approach B (center) when the interaction matrix satisfies condition (5.27) for the projection distance. Right: Kinetic energy.

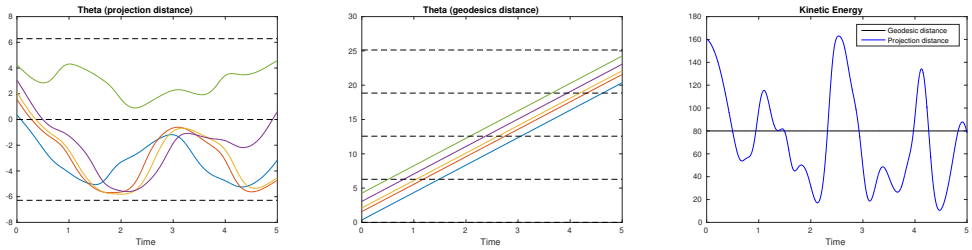


Figure 5.6: Evolution of the system (5.26) with Approach A (left) Approach B (center) when the interaction matrix satisfies condition (5.27) for the geodesic distance. Right: Kinetic energy.

5.3 Analysis and simulations on $\mathbb{S}^2$

5.3.1 Models

We study both approaches A and B for $M = \mathbb{S}^2$, i.e. for a two dimensional sphere embedded in $\mathbb{R}^3$. We use spherical coordinates: let $(\theta_i)_{i\in\{1,\dots,N\}} \in [0, 2\pi]^N$, and $(\phi_i)_{i\in\{1,\dots,N\}} \in [0, \pi]^N$ such that for all $i \in \{1, ..., N\}, x_i = (\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi)^T$.

Choice of influence function We choose an influence function $\Psi(d)$ between two agents x_i and x_j so that the right-hand side of the system is continuous, the discontinuities are shown in Table 5.1. For approach B, the only point in $\mathcal{CL}(x_i)$ for a given x_i is the antipodal point (this is an end point of a diameter for which x_i is the other end point.) As in the case of $\mathbb{S}^1$, for approach B, we choose a function Ψ from a family of functions of the form Ψ^a , see equation (5.25).

Approach A. On $\mathbb{S}^2$, the derivative for system (5.1)-(5.2)-(5.3) with $\Psi \equiv \mathrm{Id}$ reduces to the sum of all projections onto the tangent space of agent x_i, weighted by the corresponding interaction term a_{ij}. This is rewritten as:

$$\Pi_{T_{x_i}} \sum_{j=1}^{N} a_{ij}(x_j - x_i) = \Pi_{T_{x_i}} \sum_{j=1}^{N} a_{ij}(x_j) = \sum_{j=1}^{N} a_{ij}(x_j - \langle x_j, x_i\rangle x_i)$$
$$= \sum_{j=1}^{N} a_{ij} \begin{pmatrix} \cos\theta_j \sin\phi_j \\ \sin\theta_j \sin\phi_j \\ \cos\phi_j \end{pmatrix} - \left\langle \begin{pmatrix} \cos\theta_j \sin\phi_j \\ \sin\theta_j \sin\phi_j \\ \cos\phi_j \end{pmatrix}, \begin{pmatrix} \cos\theta_i \sin\phi_i \\ \sin\theta_i \sin\phi_i \\ \cos\phi_i \end{pmatrix} \right\rangle \begin{pmatrix} \cos\theta_i \sin\phi_i \\ \sin\theta_i \sin\phi_i \\ \cos\phi_i \end{pmatrix}.$$

Approach B. The geodesic distance $d_G(x_i, x_j)$ from (5.5) between two points x_i, and x_j on $\mathbb{S}^2$ is given by:

$$d_G(x_i, x_j) = 2 \arcsin\left(\frac{\|x_i - x_j\|}{2}\right),$$

and the direction toward x_j from x_i is

$$\nu_{ij} = \frac{x_j - \langle x_j, x_i \rangle x_i}{\|x_j - \langle x_j, x_i \rangle x_i\|},$$

where $\|\cdot\|$ is the standard norm in $\mathbb{R}^3$.

5.3.2 Example

Example 5.3.1. *To assess the influence of the curvature of $\mathbb{S}^2$ on the dynamics, observe a simple case involving 3 agents evolving according to the interaction matrix:*

$$A = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \tag{5.28}$$

In Section 5.5.2, we prove that those dynamics in $\mathbb{R}^2$ lead to periodic trajectories on a single orbit shared by all three agents, the orbit's parameters being fully determined by the initial conditions (see Theorem 5.5.2). However, the same dynamics on the sphere do not give rise to periodic trajectories. In sections 5.4.3, we also discuss the dynamics with this interactions matrix on $\mathbb{T}^2$, to assess the effect of curvature of the manifold.

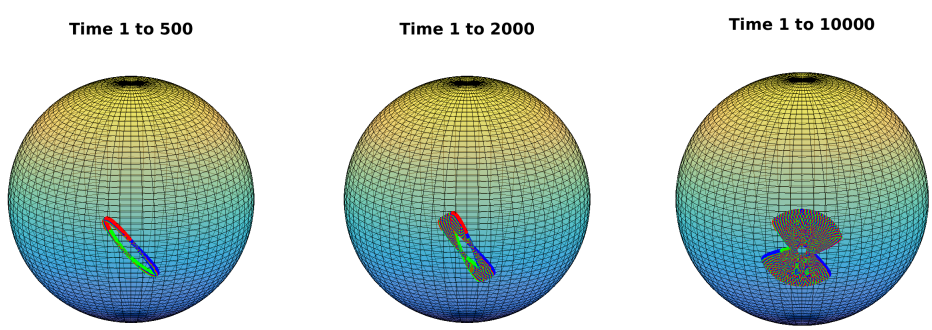


Figure 5.7: Dynamics with approach A on $\mathbb{S}^2$, using the interactions matrix 5.28. If the agents' initial positions are close enough to each other, the agents with will form trajectories that remain in a neighborhood of their initial position.

5.4 Analysis and simulations on $\mathbb{T}^2$

We now study how the general dynamics given by equation (5.1) apply to the specific case of the torus $\mathbb{T}^2 \subset \mathbb{R}^3$. Let (e_x, e_y, e_z) denote the Euclidean basis of $\mathbb{R}^3$. Let $(R, r) \in (\mathbb{R}^+)^2$, with $R > r$. We define the manifold $\mathbb{T}^2$ as the torus obtained by rotating the circle $(x-R)^2 + z^2 = r^2$ around the z-axis. Hence $\mathbb{T}^2$ is defined by the equation $(R - \sqrt{x^2+y^2})^2 + z^2 = r^2$. The parametric equations for such a torus are:

$$\begin{cases} x = (R + r\cos\theta)\cos\phi \\ y = (R + r\cos\theta)\sin\phi \\ z = r\sin\theta \end{cases} \quad \text{for } (\phi, \theta) \in [0, 2\pi)^2.$$

The angles ϕ and θ are respectively referred to as the toroidal and poloidal angles. A set of points with the same toroidal angle is called a meridian.

5.4.1 Model

We first investigate the behavior of system (5.1) with approach B (using the geodesic distance) in the case of $\mathbb{T}^2$. Unlike in the cases of $\mathbb{S}^1$ and $\mathbb{S}^2$ presented in the sections 5.2 and 5.3, there exists no simple expression for the geodesic distance between two points on the torus. In 1903, Bliss studied and classified the different kinds of geodesic lines on the standard torus [11], using elliptic functions. Gravesen et al. determined the structure of the cut loci of a torus of revolution [46].

Several challenges arise when defining approach B on $\mathbb{T}^2$. Firstly, computing the Riemannian distance between two points is highly non-trivial. One could consider approximating it numerically, but in the numerical discretization of equations (5.1)-(5.5)-(5.6), $N(N-1)/2$ geodesics would have to be computed per time-step. That would require tremendous computing power.

Secondly, assuming that one is able to efficiently compute the geodesics on $\mathbb{T}^2$, one must take into account the cut-loci of each point to ensure that the dynamics (5.1)-(5.5)-(5.6) are well-defined. A method to guarantee well-defined dynamics would be to use a bounded confidence model [55], where the neighborhood of influence for an agent x_i at point p is of smaller radius than the closest element in the cut locus of p. See section 5.1.2 for conditions on Ψ to make the right hand side of equation (5.1) continuous.

For simplicity, we thus focus on Approach A, where the dynamics are a function of the projection of each vector $x_j - x_i$ onto the tangent space at x_i. We will show that some restrictions still apply to the interaction function Ψ, but they are less restrictive and more easily determined than in Approach

B.

Equations (5.1)-(5.2) reads:

$$\dot{x}_i = \sum_{j=1}^{N} a_{ij}\Psi(\|\Pi_{T_{x_i}\mathbb{T}^2}(x_j - x_i)\|)\nu_{ij}, \quad i \in \{1, \dots, N\}. \tag{5.29}$$

The vector ν_{ij} depends on the influence that x_j has over x_i. It is zero if $\Pi_{T_{x_i}\mathbb{T}^2}(x_j - x_i) = 0$, and it is a unit vector otherwise. Let $\mathcal{N}_i$ be the set of points that have no influence on x_i (see Table 5.1). Then, given $i, j \in \{1, \dots, N\}$, ν_{ij} has the following expression:

$$\nu_{ij} = \begin{cases} \frac{\Pi_{T_{x_i}\mathbb{T}^2}(x_j - x_i)}{\|\Pi_{T_{x_i}M}(x_j - x_i)\|} & \text{if } x_j \notin \mathcal{N}_i \\ 0 & \text{if } x_j \in \mathcal{N}_i. \end{cases} \tag{5.30}$$

Let $x_i \in \mathbb{T}^2$. We start by determining the set $\mathcal{N}_i$. For all i, we define the vectors $u_{\phi_i} = \cos\phi_i e_x + \sin\phi_i e_y$ and $u_{\theta_i} = \cos\theta_i u_{\phi_i} + \sin\theta_i e_z$, so that each agent's position vector reads: $x_i = Ru_{\phi_i} + ru_{\theta_i}$. With these notations, u_{θ_i} is the normal to the tangent space at the point x_i. A basis for the tangent space at a point $x_i(\phi_i, \theta_i)$ is given by the two tangent vectors $t_{\phi_i} = (-\sin\phi_i, \cos\phi_i, 0)$ and $t_{\theta_i} = (-\sin\theta_i\cos\phi_i, -\sin\theta_i\sin\phi_i, \cos\theta_i)$. Notice that $\langle x_i, t_{\phi_i}\rangle = 0$. Hence the condition $\Pi_{T_{x_i}\mathbb{T}^2}(x_j - x_i) = 0$ reads:

$$\begin{cases} \langle x_j, t_{\phi_i}\rangle = 0 \\ \langle x_j - x_i, t_{\theta_i}\rangle = 0. \end{cases}$$

After computations, we get:

$$\langle x_j, t_{\phi_i}\rangle = 0 \iff \sin(\phi_j - \phi_i) = 0 \iff \phi_j = \phi_i + k\pi, k \in \mathbb{Z}.$$

If $\phi_j = \phi_i$, the second condition becomes:

$$\langle x_j - x_i, t_{\theta_i}\rangle = 0 \iff \sin(\theta_j - \theta_i) = 0 \iff \theta_j = \theta_i + k\pi, k \in \mathbb{Z}.$$

If $\phi_j = \phi_i \pm \pi$, the second condition becomes:

$$\sin(\theta_i + \theta_j) = -\frac{2R}{r}\sin\theta_i.$$

Notice that this last equation only has a solution if $|\sin\theta_i| \leq \frac{r}{2R}$. The set of positions that have no influence on x_i thus comprises up to four points on the torus, depending on the values of r, R and

$\sin\theta_i$. We then have: $\mathcal{N}_i = \{(\phi_i, \theta_i),\ (\phi_i, -\theta_i),\ (-\phi_i, -\theta_i - \mathrm{sgn}(\sin\theta_i)\arcsin(|\frac{2R}{r}\sin\theta_i|), (-\phi_i, \pi - \theta_i + \mathrm{sgn}(\sin\theta_i)\arcsin(|\frac{2R}{r}\sin\theta_i|))\}$. To ensure the continuity of the right-hand side of equation (5.29), one must impose the conditions of table 5.1.

We now go back to equation (5.29). We study the specific case where $\Psi \equiv \mathrm{Id}$, which indeed satisfies (5.1). Then the system becomes:

$$\dot{x}_i = \Pi_{T_{x_i}\mathbb{T}^2}\left(\sum_{j=1}^{N} a_{ij}(x_j - x_i)\right). \tag{5.31}$$

Hence the velocity reads:

$$\dot{x}_i = \sum_{j=1}^{N} a_{ij}(x_j - x_i) - \langle \sum_{j=1}^{N} a_{ij}(x_j - x_i), u_{\theta_i}\rangle u_{\theta_i} = \alpha_i - \langle \alpha_i, u_{\theta_i}\rangle u_{\theta_i} - \left(\sum_{j=1}^{N} a_{ij}\right)\langle x_i, t_{\theta_i}\rangle t_{\theta_i}$$

where $\alpha_i := \sum_{j=1}^{N} a_{ij}x_j$ is the sum of the influences of all agents on agent i. Notice that with the same notation, the system does not reduce to the simple form $\dot{x}_i = \alpha_i - \langle \alpha_i, x_i\rangle x_i$ for the same dynamics on the sphere (see [19]). This is due to the fact that on the torus, the position vector x_i does not define the normal to the tangent space at x_i, unlike in the cases of $\mathbb{S}^1$ and $\mathbb{S}^2$.

The velocity of each agent is given by:

$$\dot{x}_i = \begin{pmatrix} -\dot{\phi}_i \sin\phi_i(R + r\cos\theta_i) - r\dot{\theta}_i \sin\theta_i \cos\phi_i \\ \dot{\phi}_i \cos\phi_i(R + r\cos\theta_i) - r\dot{\theta}_i \sin\theta_i \sin\phi_i \\ r\dot{\theta}_i \cos\theta_i \end{pmatrix} = \dot{\phi}_i(R + r\cos\theta_i)t_{\phi_i} + r\dot{\theta}_i t_{\theta_i}. \tag{5.32}$$

From (5.31) and (5.32) we get the angular velocities:

$$\begin{cases} \dot{\phi}_i = \frac{1}{(R + r\cos\theta_i)}\langle \sum_{j=1}^{N} a_{ij}(x_j - x_i), t_{\phi_i}\rangle \\ \dot{\theta}_i = \frac{1}{r}\langle \sum_{j=1}^{N} a_{ij}(x_j - x_i), t_{\theta_i}\rangle. \end{cases} \tag{5.33}$$

Notice that unlike in the case of $\mathbb{S}^2$, here the derivatives $\dot{\phi}_i$ and $\dot{\theta}_i$ are not singular.

5.4.2 Properties

We now analyze the dynamics (5.1)-(5.2)-(5.3) on $\mathbb{T}^2$. We identify families of initial conditions that trivialize the dynamics.

Proposition 5.4.1. *Consider the dynamics* (5.1)-(5.2)-(5.3) *on* $M = \mathbb{T}^2$. *Let* $P_z := \{(x, y, z) \in$

$\mathbb{R}^3 \mid z = 0\}$. *Let* $x_i(t)$ *be the position of the ithe agent at time* t. *If for all* $i \in \{1, \ldots, N\}$, $x_i(0) \in \mathbb{T}^2 \cap P_z$, *then for all* $t \geq 0$, *for all* $i \in \{1, \ldots, N\}$, $x_i(t) \in \mathbb{T}^2 \cap P_z$.

Proof. Suppose that for all $i \in \{1, \ldots, N\}$, $x_i(0) \in \mathbb{T}^2 \cap P_z$. Then for all $i \in \{1, \ldots, N\}$, $\theta_i(0) = 0$ or $\theta_i(0) = \pi$. Hence, for all $i, j \in \{1, \ldots, N\}$,

$$t_{\theta_i}(0) = \begin{pmatrix} 0 \\ 0 \\ \pm\pi \end{pmatrix} \quad \text{and} \quad x_j(0) - x_i(0) = \begin{pmatrix} (R + r\cos\theta_j)\cos\phi_j - (R + r\cos\theta_i)\cos\phi_i \\ (R + r\cos\theta_j)\sin\phi_j - (R + r\cos\theta_i)\sin\phi_i \\ 0 \end{pmatrix}$$

From equation (5.33) we get: for all $i \in \{1, \ldots, N\}$, $\dot{\theta}_i = 0$. By uniqueness of solution, for all $i \in \{1, \ldots, N\}$, $\theta_i(t) = \theta_i(0)$. All the initial velocities belong to the plane P_z. Hence all agents remain on P_z at all time. □

Remark 5.4.1. *As a consequence of Proposition 5.4.1, if all agents are initially in* $\mathbb{T}^2 \cap P_z$, *all agents initially on the bigger circle* $\theta = 0$ *remain on the major circle at all time and all agents on the minor circle* $\theta = \pi$ *remain on the minor circle at all time. In particular, if all agents are initially all on the same circle (i.e. for all* $i \in \{1, \ldots, N\}$, $\theta_i = 0$ *or for all* $i \in \{1, \ldots, N\}$, $\theta_i = \pi$*), then the torus dynamics simplify to the dynamics on* $\mathbb{S}^1$ *given by* (5.21) *or* (5.22).

Proposition 5.4.2. *Consider the dynamics* (5.1)-(5.2)-(5.3) *on* $M = \mathbb{T}^2$. *Let* $\tilde{\phi} \in [0, 2\pi]$ *and let* $P_{\tilde{\phi}} := \{(x, y, z) \in \mathbb{R}^3 \mid y = \tan(\tilde{\phi})x\}$. *If for all* $i \in \{1, \ldots, N\}$, $x_i(0) \in \mathbb{T}^2 \cap P_{\tilde{\phi}}$, *then for all* $t \geq 0$, *for all* $i \in \{1, \ldots, N\}$, $x_i(t) \in \mathbb{T}^2 \cap P_{\tilde{\phi}}$.

Proof. Suppose without loss of generality that $\bar{\phi} = 0$. Similarly to the proof for Proposition 5.4.1, we can show that for all $i \in \{1, \ldots, N\}$, $\dot{\phi}_i(0) = 0$. By uniqueness of solution, for all $i \in \{1, \ldots, N\}$, $\phi_i(t) = \phi_i(0)$. Hence all agents remain in $P_{\tilde{\phi}}$ at all time. □

Remark 5.4.2. *As a consequence of Proposition 5.4.2, if all agents are initially in* $\mathbb{T}^2 \cap P_{\bar{\phi}}$, *all agents initially on the circle* $\phi = \bar{\phi}$ *remain on that circle at all time and all agents on the circle* $\phi = -\bar{\phi}$ *remain on that circle at all time. In particular, if all agents are initially all on the same circle (i.e. for all* $i \in \{1, \ldots, N\}$, $\phi_i = \bar{\phi}$ *or for all* $i \in \{1, \ldots, N\}$, $\phi_i = -\bar{\phi}$*), then the torus dynamics simplify to the dynamics on* $\mathbb{S}^1$ *given by* (5.21) *or* (5.22).

5.4.3 Simulations

To assess the influence of the curvature of the manifold on the dynamics, we compare a simple case involving 3 agents evolving according to the interaction matrix given in equation (5.28). As in the

case of $\mathbb{S}^2$, the dynamics on the torus do not give rise to periodic trajectories (as opposed to the dynamics in $\mathbb{R}^2$, see Theorem 5.5.2). Instead, since $\mathbb{T}^2$ can locally be identified with $\mathbb{R}^2$, if the initial mutual distances are small enough, the dynamics resemble those in $\mathbb{R}^2$. More specifically, the trajectories are quasi-periodic with a gradual shift of the center of mass (see Figure 5.8). However, if the initial distances between agents are large, the geometry and curvature of the torus changes radically the behavior of the system.

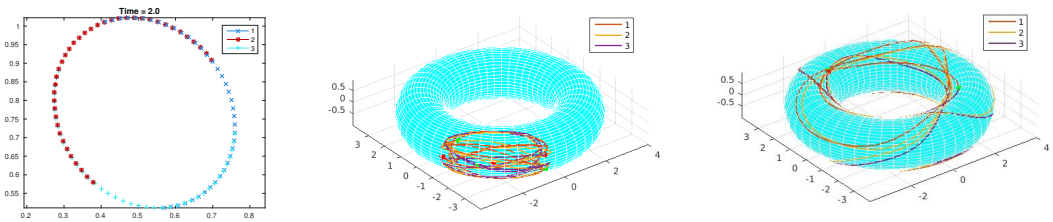


Figure 5.8: Trajectories of three agents interacting according to the matrix A given in (5.28). Left: Dynamics in $\mathbb{R}^2$, with periodic trajectories on a unique orbit. Center: Dynamics on $M = \mathbb{T}^2$ with small initial mutual distances. Right: Dynamics on $M = \mathbb{T}^2$ with large initial distances.

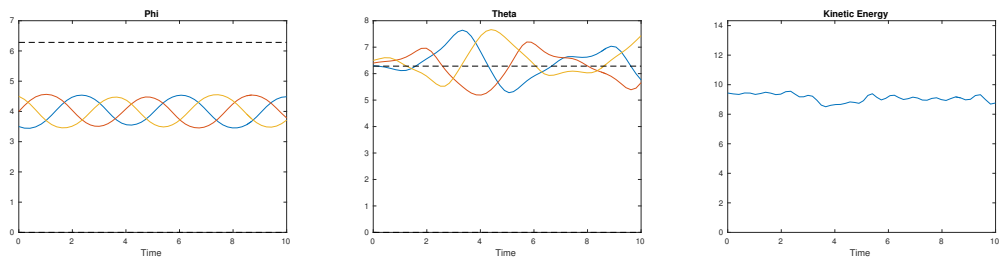


Figure 5.9: Evolutions of the coordinates of the three agents evolving on $\mathbb{T}^2$ with interaction matrix A from equation (5.28), with small initial mutual distances. Left: Evolution of ϕ. Center: Evolution of θ. Right: Evolution of the kinetic energy.

5.5 Social choreographies

As seen in Sections 5.2 and 5.4.3, when the interaction matrix A satisfies certain properties, for instance given by (5.27) on $\mathbb{S}^1$ or by (5.28) in $\mathbb{R}^2$, then the trajectories exhibit special properties of symmetry or periodicity. In [19], configurations on $\mathbb{S}^2$ in which all mutual distances between agents remain constant were named *dancing equilibrium*.

In this section, we investigate systems with similar properties of periodicity or symmetry. We use the term *social choreography*, drawing a parallel with the well-known "n-body choreographies" discovered by Moore [80, 81] in the context of point masses subject to gravitational forces. In the n-body problem, the interaction potentials between masses are predetermined, as they depend exclu-

sively on the masses and distances between agents. Hence the conditions for a n-body choreography to occur only depend on the initial state of the system. In the case of social choreography, there are more degrees of freedom, as we design the interaction matrix as well as to set the initial conditions. We study sufficient conditions on the interaction matrices for the trajectories of the system to be periodic or symmetric, focusing on the Euclidean space $\mathbb{R}^2$ with the specific choice of interaction potential $\Psi \equiv \mathrm{Id}$. In this setting, both approaches A and B are equivalent and the system simply reads as:

$$\text{for all } i \in \{1, \dots, N\}, \quad \dot{x}_i = \sum_{j=1}^{N} a_{ij}(x_j - x_i). \tag{5.34}$$

A simple case of social choreography is that of a system with periodic trajectories, which we define as follows:

Definition 5.5.1. *Let $(x_i)_{i=1\dots N}$ be a solution of (5.34). We refer to the system as having* ***periodic trajectories*** *if there exists $\tau > 0$ such that*

$$\textit{for all } i \in \{1, ..., N\}, \textit{ for all } t > 0, \quad x_i(t+\tau) = x_i(t).$$

We will examine possible periodic behaviors of the system in sections 5.5.2, 5.5.3 and 5.5.4.

5.5.1 Rotationally invariant system

We now give sufficient conditions on the interaction matrix and on the initial conditions for the system to be invariant by rotation.

Theorem 5.5.1. *Let $k \in \mathbb{N}$ such that k divides N. Let $P_k = \begin{pmatrix} 0 & I_{N-k} \\ I_k & 0 \end{pmatrix}$ be the matrix of change of basis from $(e_1, \dots, e_N)$ to $(e_k, \dots, e_N, e_1, \dots, e_{k-1})$. Let $R(\theta)$ denote the rotation matrix in $\mathbb{R}^2$ for the angle $\theta \in [0, 2\pi)$. Suppose that initially, the system is invariant by rotation of angle $\frac{2k\pi}{N}$, that is:*

$$\textit{for all } i \in \{1, \dots, N\}, \quad R(\frac{2k\pi}{N})x_i(0) = \begin{cases} x_{i+k}(0) & \text{if } i+k \leq N \\ x_{i+k-N}(0) & \text{if } i+k > N \end{cases}.$$

Suppose that the interaction matrix A is invariant by change of basis, i.e. $P_k^{-1}AP_k = A$. Then the system remains invariant by rotation of angle $\frac{2k\pi}{N}$ at all time:

$$\textit{for all } t > 0, \quad \textit{for all } i \in \{1, \dots, N\}, \quad R(\frac{2k\pi}{N})x_i(t) = \begin{cases} x_{i+k}(t) & \text{if } i+k \leq N \\ x_{i+k-N}(t) & \text{if } i+k > N \end{cases}.$$

Proof. Let $A \in \mathcal{M}^N(\mathbb{R})$ be the interaction matrix, i.e. $A = (a_{ij})_{i,j=1,\dots N}$, and define $D = \mathrm{diag}(\sum_j a_{ij})$. Let $x = (x_1, \dots, x_N)$ denote the set of all x_i's. It is a vector of length N with entries in $\mathbb{R}^2$. Let $X \in \mathcal{M}^{N\times 2}(\mathbb{R})$ denote the corresponding matrix of $\mathbb{R}^{N\times 2}$ such that for all $i \in \{1, \dots, N\}$, for all $j \in \{1, 2\}$, X_{ij} is the j-th coordinate of x_i. With these notations, $\dot{X} = \tilde{A}X$, where $\tilde{A} = A - D$. We denote by $(e_1, \dots, e_N)$ the canonical orthonormal basis of $(\mathbb{R})^N$ such that $X = \sum_{i=1}^N e_i x_i^T$.

From the definition of the matrix X, the condition

$$\text{for all } i \in \{1, \dots, N\}, \quad R(\frac{2k\pi}{N})x_i(0) = \begin{cases} x_{i+k}(0) & \text{if } i+k \leq N \\ x_{i+k-N}(0) & \text{if } i+k > N \end{cases}$$

can be rewritten as: $P_k X(0) = (R(\frac{2k\pi}{N})X(0)^T)^T$. Let $Y := P_k X$ and $Z := (R(\frac{2k\pi}{N})X^T)^T$. From the theorem's hypotheses, $Y(0) = Z(0)$. Let us show that Y and Z have the same evolution. One can easily prove that $P_k^{-1}\tilde{A}P_k$ if and only if $P_k^{-1}AP_k$. Then notice that

$$\dot{X} = \tilde{A}X = P_k^{-1}\tilde{A}P_k X.$$

From that we compute:

$$\dot{Y} = P_k\dot{X} = P_k(P_k^{-1}\tilde{A}P_k X) = \tilde{A}P_k X = \tilde{A}Y.$$

Similarly,

$$\dot{Z} = (R(\frac{2k\pi}{N})\dot{X}^T)^T = (R(\frac{2k\pi}{N})(\tilde{A}X)^T)^T = (R(\frac{2k\pi}{N})X^T\tilde{A}^T)^T = \tilde{A}Z.$$

Since Y and Z satisfy the same differential equation and $Y(0) = Z(0)$, then $Y(t) = Z(t)$ for all $t \geq 0$. This implies that at all time,

$$\text{for all } i \in \{1, \dots, N\}, \quad R(\frac{2k\pi}{N})x_i(t) = \begin{cases} x_{i+k}(t) & \text{if } i+k \leq N \\ x_{i+k-N}(t) & \text{if } i+k > N \end{cases}.$$

□

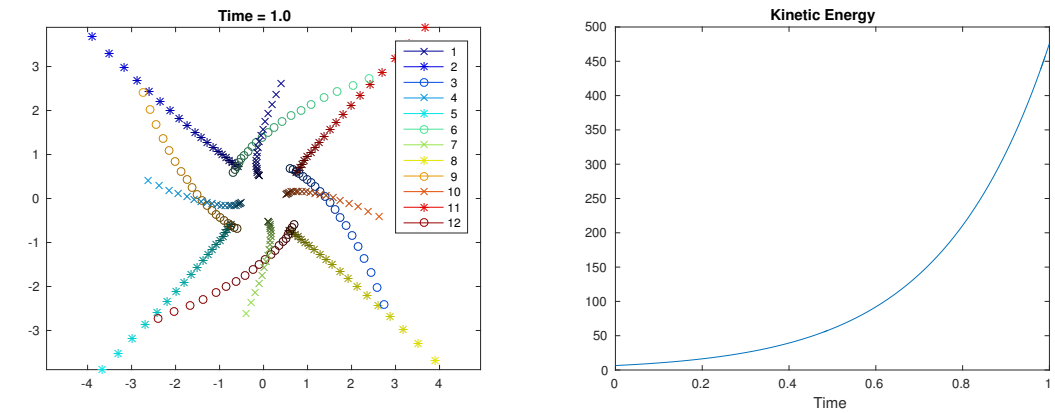


Figure 5.10: Left: Evolution of 12 agents with the conditions of Theorem 5.5.1, with $k = 3$, resulting in diverging trajectories. Dark to light color scale indicates earlier to later time. Right: corresponding exploding kinetic energy. The interaction matrix A and the initial positions were generated according to a random algorithm, with the conditions of Theorem 5.5.1.

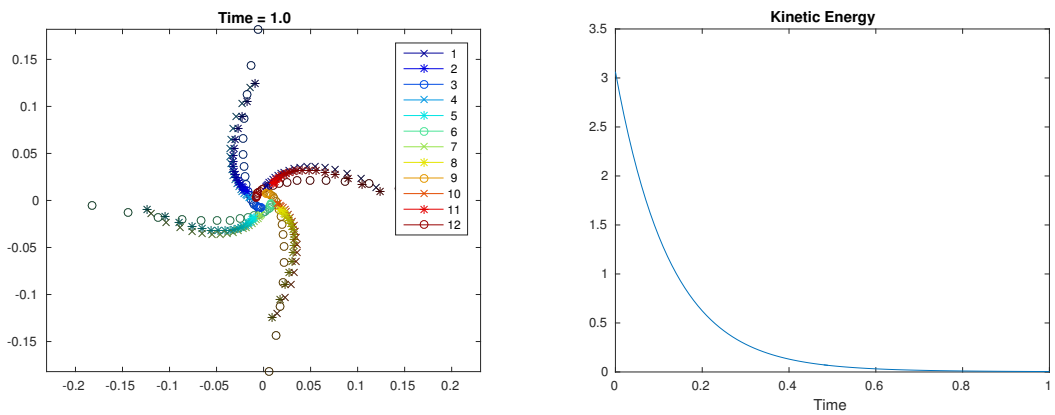


Figure 5.11: Left: Evolution of 12 agents with the conditions of Theorem 5.5.1, with $k = 3$, resulting in convergence to consensus. Dark to light color scale indicates earlier to later time. Right: corresponding kinetic energy converging to zero. The interaction matrix A and the initial positions were generated according to a random algorithm, with the conditions of Theorem 5.5.1.

5.5.2 Unique orbit

Another example of social choreography is that of a system in which all agents share one unique orbit. Such choreographies have been discovered in the context of the n-body problem, for instance the "figure 8" orbit for three equal masses [80].

Definition 5.5.2. *Let $(x_i)_{i=1\dots N}$ be a solution of* (5.34)*. We say that the system has a* ***unique orbit*** *if the orbits of all points are identical, i.e.*

$$\textit{for all } i, j \in \{1, ..., N\}, \qquad \{z \in M | \exists t > 0,\ x_i(t) = z\} = \{z \in M | \exists t > 0,\ x_j(t) = z\}.$$

To illustrate Theorem 5.5.1, we study the evolution of N agents initially positioned at regular intervals on a circle, with an interaction matrix and initial conditions given by:

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & -1 \\ -1 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 1 \\ 1 & 0 & \dots & 0 & -1 & 0 \end{pmatrix} \quad \text{and} \quad \text{for all } i \in \{1, \dots, N\}, \quad x_i(0) = \begin{pmatrix} \cos(\frac{2i\pi}{N}) \\ \sin(\frac{2i\pi}{N}) \end{pmatrix}. \tag{5.35}$$

Notice that $\tilde{A} = A$, and the system satisfies the conditions of Theorem 5.5.1 with $k = 1$. Hence for all $i \in \{1, \dots, N-1\}$, $R(\frac{2\pi}{N})x_i(t) = x_{i+1}(t)$ and $R(\frac{2\pi}{N})x_N(t) = x_1(t)$. The $2N$-dimensional system then reduces to a 2-dimensional one for the two coordinates x_{11} and x_{12} of x_1, and all the other variables can be recovered by rotation of x_1:

$$\dot{x}_1 = x_2 - x_N = R(\frac{2\pi}{N})x_1 - R(-\frac{2\pi}{N})x_1.$$

This can be written as:

$$\begin{pmatrix} \dot{x}_{11} \\ \dot{x}_{12} \end{pmatrix} = \begin{pmatrix} 0 & -2\sin(\frac{2\pi}{N}) \\ 2\sin(\frac{2\pi}{N}) & 0 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}.$$

Solving this linear system yields:

$$\begin{cases} x_{11}(t) = x_{11}(0)\cos(2\sin(\frac{2\pi}{N})t) - x_{12}(0)\sin(2\sin(\frac{2\pi}{N})t) = \cos(2\sin(\frac{2\pi}{N})t) \\ x_{12}(t) = x_{11}(0)\sin(2\sin(\frac{2\pi}{N})t) + x_{12}(0)\cos(2\sin(\frac{2\pi}{N})t) = \sin(2\sin(\frac{2\pi}{N})t) \end{cases}.$$

This proves that all agents share one common circular orbit, and their trajectories are periodic of period $2\pi(2\sin(\frac{2\pi}{N}))^{-1}$. Figure 5.12 provides a numerical illustration of this behavior, with 10 agents initially positioned at regular intervals on the unit circle.

Another interesting example is that of 3 agents interacting according to the interaction matrix

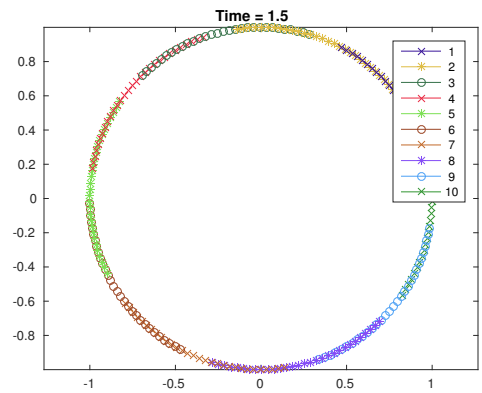


Figure 5.12: Periodic trajectories of 10 agents sharing one circular orbit

given previously, which, reduced to $N = 3$, gives:

$$A = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}. \tag{5.36}$$

Theorem 5.5.2. *Let $N = 3$. Consider the system* (5.34) *with interaction matrix given by* (5.36)*. Then there exists a unique orbit shared by all agents, and all three trajectories are periodic.*

Proof. The x and y-coordinates of the systems are decoupled, so that the 6-dimensional system can be reduced to two 3-dimensional ones. Notice that $\tilde{A} = A$. Then for each coordinate $j \in \{1, 2\}$, the system reads:

$$\begin{pmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \end{pmatrix}(t) = \exp(tA) \begin{pmatrix} x_{1j}^0 \\ x_{2j}^0 \\ x_{3j}^0 \end{pmatrix}$$

with

$$e^{tA} = \frac{1}{3} \begin{pmatrix} 1 + 2\cos(\sqrt{3}t) & 1 - \cos(\sqrt{3}t) + \sqrt{3}\sin(\sqrt{3}t) & 1 - \cos(\sqrt{3}t) - \sqrt{3}\sin(\sqrt{3}t) \\ 1 - \cos(\sqrt{3}t) + \sqrt{3}\sin(\sqrt{3}t) & 1 - \cos(\sqrt{3}t) - \sqrt{3}\sin(\sqrt{3}t) & 1 + 2\cos(\sqrt{3}t) \\ 1 - \cos(\sqrt{3}t) - \sqrt{3}\sin(\sqrt{3}t) & 1 + 2\cos(\sqrt{3}t) & 1 - \cos(\sqrt{3}t) + \sqrt{3}\sin(\sqrt{3}t) \end{pmatrix}.$$

Due to the special structure of e^{tA}, this can be rewritten as:

$$\begin{pmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \end{pmatrix}(t) = \frac{1}{3}\begin{pmatrix} x_{1j}^0 & x_{2j}^0 & x_{3j}^0 \\ x_{2j}^0 & x_{3j}^0 & x_{1j}^0 \\ x_{3j}^0 & x_{1j}^0 & x_{2j}^0 \end{pmatrix}\begin{pmatrix} 1+2\cos(\sqrt{3}t) \\ 1-\cos(\sqrt{3}t)+\sqrt{3}\sin(\sqrt{3}t) \\ 1-\cos(\sqrt{3}t)-\sqrt{3}\sin(\sqrt{3}t) \end{pmatrix}.$$

This shows that all three trajectories are periodic, or period $\frac{2\pi}{\sqrt{3}}$. One can compute the positions of each agent after a third of a period and notice that:

$$\begin{pmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \end{pmatrix}(t+\frac{2\pi}{3\sqrt{3}}) = \frac{1}{3}\begin{pmatrix} x_{1j}^0 & x_{2j}^0 & x_{3j}^0 \\ x_{2j}^0 & x_{3j}^0 & x_{1j}^0 \\ x_{3j}^0 & x_{1j}^0 & x_{2j}^0 \end{pmatrix}\begin{pmatrix} 1-\cos(\sqrt{3}t)-\sqrt{3}\sin(\sqrt{3}t) \\ 1+2\cos(\sqrt{3}t) \\ 1-\cos(\sqrt{3}t)+\sqrt{3}\sin(\sqrt{3}t) \end{pmatrix} = \begin{pmatrix} x_{2j} \\ x_{3j} \\ x_{1j} \end{pmatrix}(t).$$

This shows that there is one unique shared orbit. □

5.5.3 Coupled periodic trajectories

Other conditions on the interaction matrix A give rise to different kinds of periodic behaviors. Here we provide sufficient conditions for the system to exhibit periodic trajectories, such that each orbit is shared by two agents.

Theorem 5.5.3 (Coupled periodic trajectories)**.** *Let N be even. Suppose that initially, the system is invariant by rotation of angle $\frac{4\pi}{N}$, that is:*

$$\textit{for all } i \in \{1,\dots,N\}, \quad R(\frac{4\pi}{N})x_i(0) = \begin{cases} x_{i+2}(0) & \text{if } i+2 \le N \\ x_{i+2-N}(0) & \text{if } i+2 > N \end{cases}.$$

Let $a, b > 0$ and let

$$A = \begin{pmatrix} 0 & a & 0 & \dots & 0 & -b \\ -a & 0 & b & \ddots & \ddots & 0 \\ 0 & -b & \ddots & a & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & a \\ b & 0 & \dots & 0 & -a & 0 \end{pmatrix}. \tag{5.37}$$

Then the system is periodic of period $\tau = \frac{\pi}{\sqrt{ab}\sin(2\pi/N)}$. Furthermore, if N is divisible by 4, opposite

agents share orbits two by two, i.e.:

$$\textit{for all } t > 0, \textit{ for all } i \in \{1, \dots, \frac{N}{2}\}, x_i(t+\tau) = x_{i+\frac{N}{2}}(t),$$

and the kinetic energy is periodic with period $\tau/2$.

Proof. First remark that the system satisfies the hypotheses of Theorem 5.5.1, so

$$\text{for all } t > 0, \quad \text{for all } i \in \{1, \dots, N\}, \quad R(\frac{4\pi}{N})x_i(t) = \begin{cases} x_{i+2}(t) & \text{if } i+2 \leq N \\ x_{i+2-N}(t) & \text{if } i+2 > N \end{cases}.$$

Hence the system is entirely known from the positions of the first two agents, since all others can be obtained by simple rotations. We show that this $2N$-dimensional problem can be rewritten as a 4-dimensional one. Indeed, using the fact that $x_N = R(-4\pi/N)x_2$ and $x_3 = R(4\pi/N)x_1$, the system

$$\begin{cases} \dot{x}_1 = a(x_2 - x_1) - b(x_N - x_1) \\ \dot{x}_2 = b(x_3 - x_2) - a(x_1 - x_2) \end{cases}$$

becomes:

$$\begin{cases} \dot{x}_1 = \begin{pmatrix} \dot{x}_{11} \\ \dot{x}_{12} \end{pmatrix} = a\left[\begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix} - \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}\right] - b\left[\begin{pmatrix} \cos(\frac{4\pi}{N}) & \sin(\frac{4\pi}{N}) \\ -\sin(\frac{4\pi}{N}) & \cos(\frac{4\pi}{N}) \end{pmatrix}\begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix} - \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}\right] \\ \dot{x}_2 = \begin{pmatrix} \dot{x}_{21} \\ \dot{x}_{22} \end{pmatrix} = b\left[\begin{pmatrix} \cos(\frac{4\pi}{N}) & -\sin(\frac{4\pi}{N}) \\ \sin(\frac{4\pi}{N}) & \cos(\frac{4\pi}{N}) \end{pmatrix}\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} - \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}\right] - a\left[\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} - \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}\right] \end{cases}.$$

This can be rewritten in matrix form as:

$$\begin{pmatrix} \dot{x}_{11} \\ \dot{x}_{12} \\ \dot{x}_{21} \\ \dot{x}_{22} \end{pmatrix} = A_4 \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{pmatrix} \tag{5.38}$$

where

$$A_4 := \begin{pmatrix} -a+b & 0 & a-b\cos(\frac{4\pi}{N}) & -b\sin(\frac{4\pi}{N}) \\ 0 & -a+b & b\sin(\frac{4\pi}{N}) & a-b\cos(\frac{4\pi}{N}) \\ -a+b\cos(\frac{4\pi}{N}) & -b\sin(\frac{4\pi}{N}) & a-b & 0 \\ b\sin(\frac{4\pi}{N}) & -a+b\cos(\frac{4\pi}{N}) & 0 & a-b \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{pmatrix}.$$

One can easily show that this reduced interaction matrix A_4 has two purely imaginary conjugate eigenvalues, $i\lambda$ and $-i\lambda$, each of multiplicity 2, where $\lambda = 2\sqrt{ab}\sin(\frac{2\pi}{N})$. Hence the solution of the system (5.38) can be written as a weighted sum of the functions $t \mapsto \cos(\lambda t)$ and $t \mapsto \sin(\lambda t)$. This implies that the system is periodic, of period

$$\tau = \frac{2\pi}{\lambda} = \frac{\pi}{\sqrt{ab}\sin(\frac{2\pi}{N})}.$$

Furthermore, if N is divisible by 4, according to Theorem 5.5.1, $x_{\frac{N}{2}+1} = -x_1$ and $x_{\frac{N}{2}+2} = -x_2$. This implies that for all $t > 0$, $x_1(t+\tau) = -x_1(t) = x_{\frac{N}{2}+1}(t)$ and $x_2(t+\tau) = -x_2(t) = x_{\frac{N}{2}+2}(t)$, so the agents x_1 and $x_{\frac{N}{2}+1}$ share an orbit, as well as all pairs of agents x_i and $x_{\frac{N}{2}+i}$ for $i \in \{1, \ldots, \frac{N}{2}\}$.

As a consequence, the kinetic energy is periodic, of period $\tau = \pi/(2\sqrt{ab}\sin(\frac{2\pi}{N}))$. If N is divisible by 4, every half period, the system is rotated by an angle π, so the kinetic energy is periodic with period $\tau/2$. □

Remark 5.5.1. *Notice that the agents sharing orbits do not interact with one another, as shown in Figure 5.13.*

An example of such a choreography is given in Figure 5.14.

Remark 5.5.2. *As a slight generalization, we provide numerical simulations illustrating a similar behavior, but with slightly different conditions: the periodic evolution of 9 agents on three distinct orbits shared three by three, see figures 5.15 and 5.16.*

5.5.4 Helical trajectories

In sections 5.5.2 and 5.5.3, we provided conditions for the trajectories of the system to be periodic. Here, we explore further the notion of periodicity by studying systems with drift, displaying helical trajectories but periodic kinetic energy.

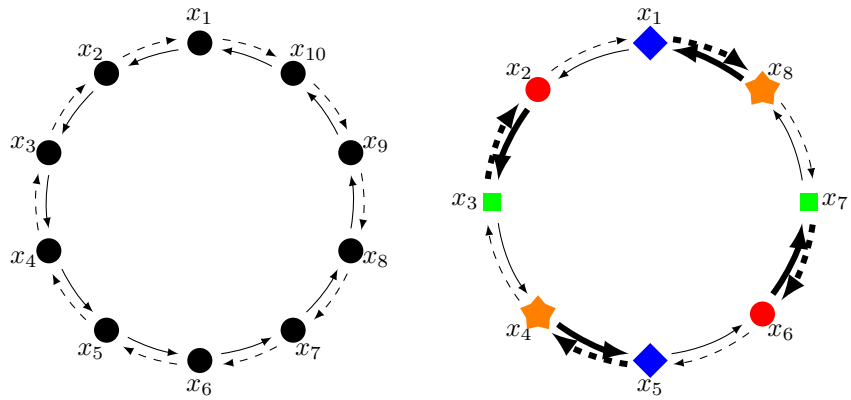


Figure 5.13: Left: Directed graph corresponding to the matrix A given in (5.35). Full arrows represent positive coefficients ($a_{ij} > 0$) while dashed ones represent negative coefficients ($a_{ij} < 0$). Right: Weighted directed graph corresponding to the matrix A given in (5.37). Thin arrows represent the weighted edges $|a_{ij}| = a$ while bold ones represent the weight $|a_{ij}| = b$. Nodes with the same color and symbol share orbits but are not directly connected in the graph.

Definition 5.5.3. *Let $(x_i)_{i=1\dots N}$ be a solution of (5.34). We call the corresponding trajectories* ***helical trajectories*** *if there exists $v \in \mathbb{R}^2$ and $\tau \in \mathbb{R}^*$ such that*

$$\textit{for all } i \in \{1, ..., N\}, \textit{ for all } t > 0, \quad x_i(t+\tau) = x_i(t) + \tau v.$$

Notice that this definition generalizes the notion of periodic trajectories recalled in Definition 5.5.1, which corresponds to the case $v = 0$. When $v \neq 0$, the system has a drift term, meaning that the relative positions between agents remain periodic but their absolute positions evolve in space.

Theorem 5.5.4. *Sufficient conditions for helical trajectories. Let $N = 4$. Let $(a, b, c, d) \in (\mathbb{R}^+)^4$ such that the interaction matrix reads*

$$A = \begin{pmatrix} 0 & a & 0 & -d \\ -a & 0 & b & 0 \\ 0 & -b & 0 & c \\ d & 0 & -c & 0 \end{pmatrix}. \tag{5.39}$$

Then the system exhibits pseudo-periodic trajectories with drift.

Proof. First notice that the first and second components x_{i1} and x_{i2} of the i-th agent's position are

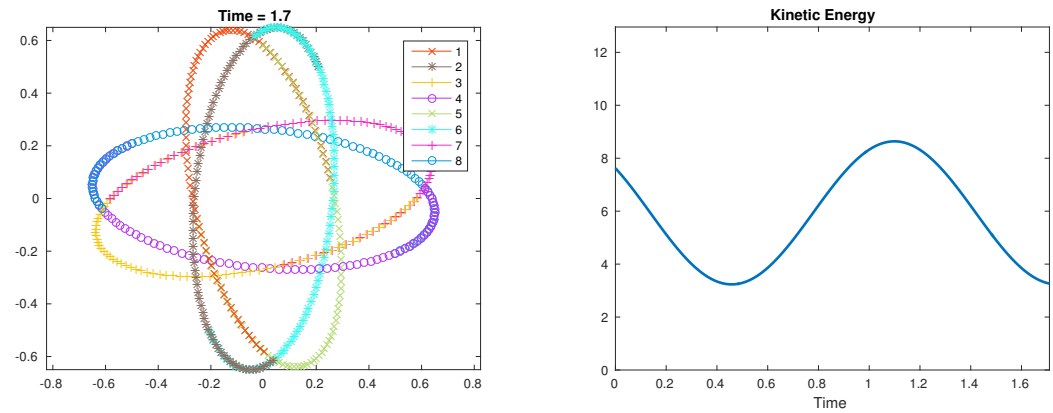


Figure 5.14: Left: Periodic trajectories of 8 agents sharing orbits two by two, in the situation of Theorem 5.5.3. Matrix A from (5.37) was constructed with $(a, b) = (1, 3)$. The initial positions $x_1(0)$ and $x_2(0)$ were randomly generated and the other 6 were obtained by rotation. The period is $\tau = 2\pi/\sqrt{6}$. Right: Corresponding kinetic energy, of period $\tau/2$.

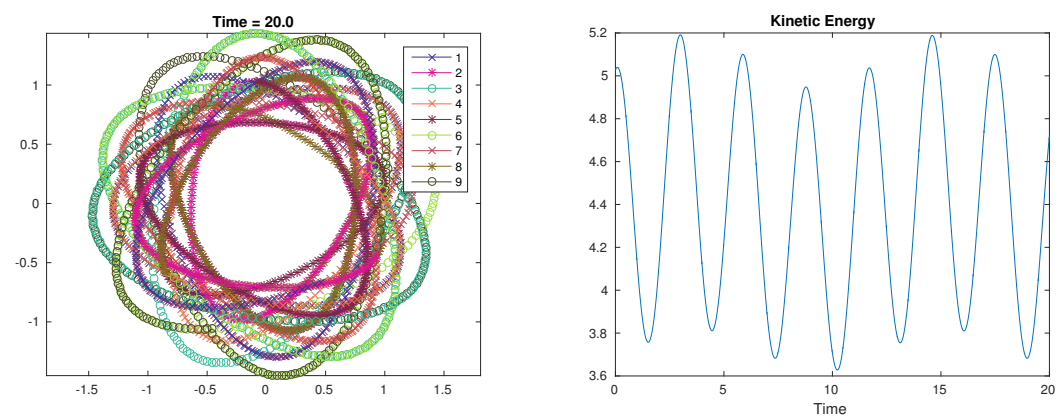


Figure 5.15: Left: evolution of 9 agents with periodic trajectories, each orbit shared by 3 agents. Right: periodic kinetic energy.

decoupled, so that the system in matrix form reads

$$\dot{x}^j = \begin{pmatrix} \dot{x}_{1j} \\ \dot{x}_{2j} \\ \dot{x}_{3j} \\ \dot{x}_{4j} \end{pmatrix} = \begin{pmatrix} d-a & a & 0 & -d \\ -a & a-b & b & 0 \\ 0 & -b & b-c & c \\ d & 0 & -c & c-d \end{pmatrix} \begin{pmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \\ x_{4j} \end{pmatrix} := \tilde{A} \begin{pmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \\ x_{4j} \end{pmatrix}, \quad \text{for } j \in \{1, 2\}. \tag{5.40}$$

Hence the projections of x on the first and second axes solve the same differential equation. The matrix $\tilde{A}$ has three distinct eigenvalues:

$$\lambda_1 = 0, \quad i\lambda_2 = i\sqrt{(a+c)(b+d)} \quad \text{and} \quad i\lambda_3 = -i\sqrt{(a+c)(b+d)}.$$

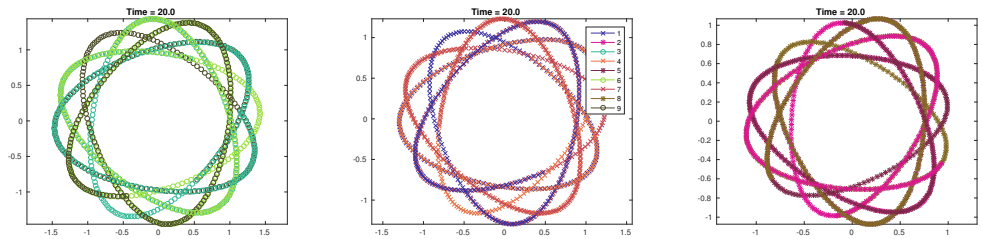


Figure 5.16: Isolated orbits of the evolution shown in Figure 5.15. Left: trajectories of agents 3, 6, 9. Middle: trajectories of agents 1, 4, 7. Right: trajectories of agents 2, 5, 8).

There is one eigenvector associated with λ_1: $v_1 := (1,1,1,1)^T$. One can show that the vectors $x(t) = v_1$ and $x(t) = v_1 t + \nu$ are both solutions of System (5.40), where, denoting $\Delta := bcd - abc + abd - acd$,

$$\nu := \frac{1}{\Delta}(ab + bc + \Delta, ab - cd + \Delta, ab + ad + \Delta, \Delta)^T.$$

Let v_2 denote the eigenvector associated with λ_2 and let v_2^R and v_2^I denote respectively its real and imaginary components, i.e. $v_2 := v_2^R + iv_2^I$. Then the solution of System (5.40) can be written as:

$$x^j(t) = C_1^j v_1 + C_2^j(v_1 t + \nu) + C_3^j \left[v_2^R \cos(\lambda_2 t) - v_2^I \sin(\lambda_2 t)\right] + C_4^j \left[v_2^R \sin(\lambda_2 t) + v_2^I \cos(\lambda_2 t)\right]$$

where $(C_1, C_2, C_3, C_4) \in \mathbb{R}^4$ are constants depending on the initial conditions. Let $\tau = \frac{2\pi}{\lambda_2}$. Then for all $t > 0$, for all $i \in \{1, \dots, 4\}$, for all $j \in \{1, 2\}$, $x_{ij}(t+\tau) = x_{ij}(t) + C_2^j \tau$. This can be rewritten as:

$$\text{for all } i \in \{1, \dots, 4\}, \text{ for all } t > 0, \quad x_i(t+\tau) = x_i(t) + \begin{pmatrix} C_2^1 \\ C_2^2 \end{pmatrix} \tau.$$

□

Theorem 5.5.5. *A system with pseudo-periodic trajectories with drift has periodic kinetic energy.*

Proof. Supose that $(x_i)_{i=1\dots N}$ has pseudo-periodic trajectories with drift, i.e. there exists $\tau \in \mathbb{R}$, $v \in \mathbb{R}^2$ such that for all $i \in \{1, \dots, N\}$, for all $t \geq 0$, $x_i(t+\tau) = x_i(t) + \tau v$. Then $\dot{x}_i(t+\tau) = \dot{x}_i(t)$ and so $E(t+\tau) = E(t)$. □

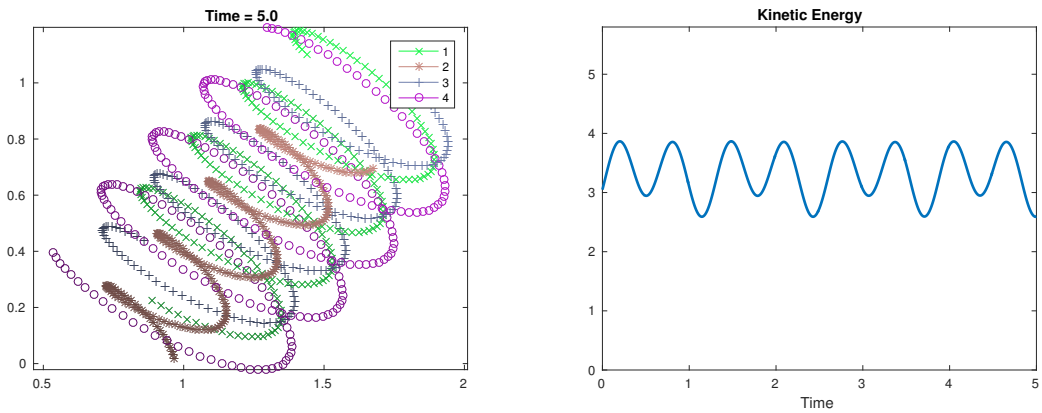


Figure 5.17: Left: Trajectories of 4 agents with helical trajectories. Parameters for matrix A (5.39) chosen to be $(a, b, c, d) = (1, 2, 3, 4)$. Dark to light color indicates earlier to later time. Right: Corresponding kinetic energy. The period is $\tau = 2\pi((a + c)(b + d))^{-1/2} = \pi/\sqrt{6}$ (see proof of Theorem 5.5.4).

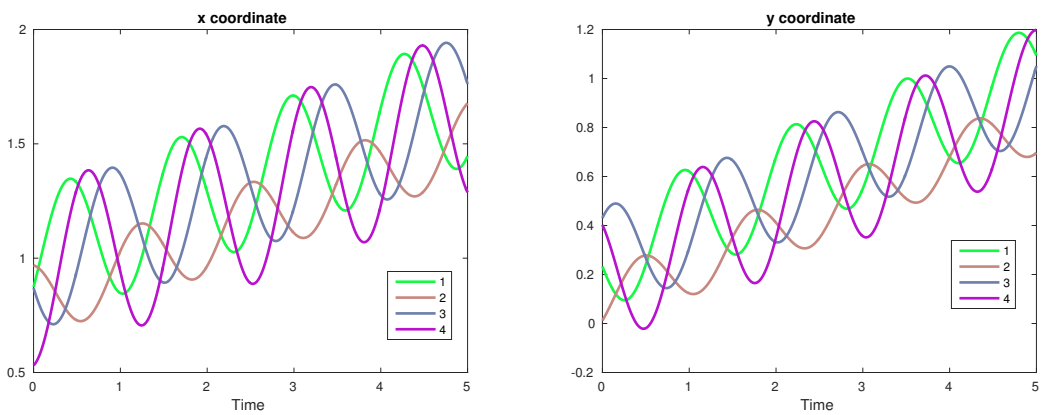


Figure 5.18: Evolution of the first and second coordinates of 4 agents with helical trajectories.

5.6 Influence of the Interaction Network

In this section we study the influence of the interaction network in bounded-confidence models. We review known properties of such models, propose open problems concerning the equilibrium sets, and provide numerical simulations illustrating the known and conjectured properties. These results were published in [3].

The Hegselmann-Krause model (HK) is a classical example of a first-order nonlinear opinion formation model [55]. It was designed in the context of opinion dynamics, and captures well-known phenomena such as formation of consensus and emergence of clustering. Agents modify their own opinion to average neighboring opinions as follows:

$$\dot{x}_i = \frac{1}{\mathbf{card}(\mathcal{S}_i)} \sum_{j \in \mathcal{S}_i} (x_j - x_i) \quad \text{for all } i \in \{1, ..., N\}, \quad x_i \in \mathbb{R}^d, \tag{5.41}$$

where $\mathcal{S}_i = \{j : \|x_i - x_j\| \leq r\}$, $r > 0$, is the set of agents interacting with agent i. The radius r can be interpreted as the level of confidence. This model captures the fact that an individual tends to trust only opinions that do not differ from its own by more than r. Since the interaction region is bounded, the HK model is also called *bounded confidence* model. Depending on the size of the interaction regions and the density of agents in the domain, different phenomena are observed. If the interaction is strong enough (i.e. r is big enough), the agents can be brought to consensus, i.e. convergence to a single opinion. If the interaction regions are too restricted, one observes clustering around different opinions. A wide variety of models have been developed by varying the confidence region $\mathcal{S}_i$. Hegselmann and Krause have for instance looked at (one-dimensional) asymmetric confidence: $\mathcal{S}_i = \{j : -r_l \leq x_i - x_j \leq r_r\}$, $r_l > 0, r_r > 0$ [55]. Recently, Motsch and Tadmor have analyzed models with interaction strength increasing with the distance between agents, showing that this so-called heterophilious dynamics enhances consensus [82].

The system can be viewed as a network represented by a (possibly time-varying) directed weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We define the set of vertices $\mathcal{V} = (\nu_i)_{i \in \{1,...,N\}}$ corresponding to the set of agents, and the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, so that an edge exists between two vertices i and j if and only if $a_{ij} \neq 0$. The edges are weighted by the interaction coefficients a_{ij}.

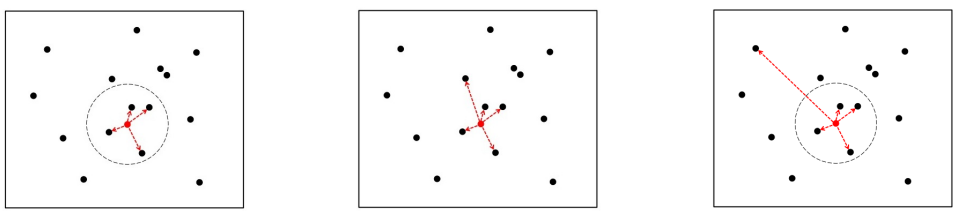
Properties of bounded-confidence models

The rationale for bounded confidence models is that it is unlikely for one agent to be influenced by another one whose opinion is too far from its own. This kind of interaction gives rise to clusters of opinions (see for instance [12]). We also mention the bounded confidence model by Deffuant, see [34] in which the opinions belong to real intervals too but the pairs of interacting agents are chosen randomly.

Two main types of interaction networks have been proposed in the literature. In metric interaction networks, agents interact depending on their distance in the state space [53]: given a confidence radius $r > 0$, we can define the interaction neighborhood $\mathcal{S}_i^r$ for the i-th agent (5.7) , see Fig. 5.19a. In topological interaction networks, agents interact depending on their relative separation. Given $k \in \mathbb{N}$, we can define the interaction neighborhood $\mathcal{S}_i^k$ (5.8), see Fig. 5.19b .

Both topological and metric interactions are local interactions. Adding long-distance connections to local ones greatly reduces the network's diameter and facilitates the spread of information [67].

This is justified by the ubiquitous idea that social networks are of small diameter, a property also known as the six degrees of separation or *small-world effect* [131]. In particular, Kleinberg (see [63, 67]) showed that single long-distance random connections in locally organized networks lead to efficient routing procedures for spreading information. The small world phenomenon is characterized by short paths (relative to the size of the network) connecting any two nodes in the network, as illustrated in Fig. 5.19c. The model as presented in [63] describes nodes on a square lattice which interact with the four adjacent nodes in the lattice, as well as one long range interaction that randomly forms an edge between a node and another non-neighboring node with a probability proportional to ρ^{-a}, where ρ is the Manhattan distance between the two nodes.



(a) Metric (b) Topological (c) Long-distance

Figure 5.19: Representation of interacting neighbors for one agent according to the different interaction networks. In (c), the long-distance connection is added to metric local interactions.

Equilibrium sets. To understand the mechanisms behind cluster formation, we studied equilibria for the HK dynamics, both with metric and topological interactions.

Let us start with metric interaction, with 2 or 3 agents in $\mathbb{R}$:

- For $N = 2$, the equilibrium set E consists of 3 subsets: The line $x_1 = x_2$; the half-plane $x_1 - x_2 > r$; the half-plane $x_2 - x_1 > r$ (see Figure 5.20a).
- In the case $N = 3$, 13 equilibrium subsets can be enumerated: the line $x_1 = x_2 = x_3$; the 3 half-planes $\{x_i = x_j,\ x_k > x_i + r\}$; the 3 half-planes $\{x_i = x_j,\ x_k < x_i - r\}$; the 6 3D manifolds $\{x_i + r < x_j < x_k - r\}$ (with i, j, k pairwise distinct in $\{1, 2, 3\}$).

Notice that in both cases, the equilibrium set is composed of pairwise disjoint manifolds with no common boundaries. We propose a general property for the equilibrium set:

Conjecture 1. *For the HK dynamics* (5.41) *with metric interaction, for all $d \in \mathbb{N}$ and $N \in \mathbb{N}$, the set of equilibria is a stratified manifold with separate strata.*

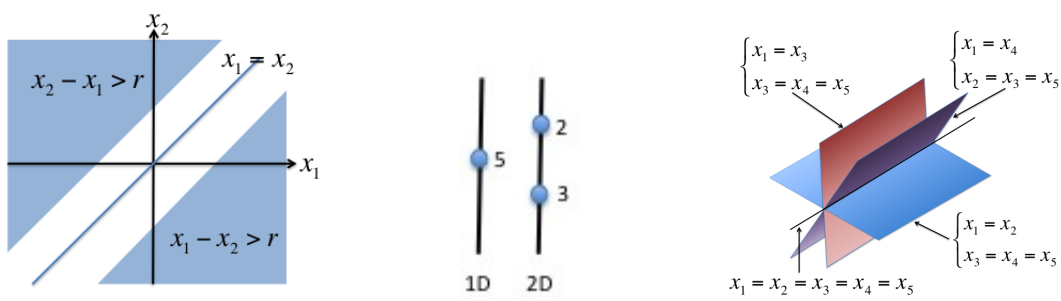
The definition of a Whitney stratified set is recalled in Def. 4.1.5. In the topological case, the number and nature of equilibrium sets depend on k.

If $k = 1$ (i.e. there is no interaction between agents), the equilibrium set is $\mathbb{R}^N$ itself. If $k = 2$ (each agent interacts with one other), we have to distinguish cases:

- for $N = 2$ or $N = 3$, the equilibrium sets are respectively the lines $x_1 = x_2$ and $x_1 = x_2 = x_3$.
- for $N \geq 4$, the equilibrium sets are more complex as they are composed of several manifolds. For instance, in the case $N = 5$, the equilibrium set consists of the line $x_1 = x_2 = x_3 = x_4 = x_5$ and the $\binom{5}{2} = 10$ half planes $\{x_i = x_j;\ x_k = x_l = x_m\}$ with i, j, k, l, m pairwise distinct in $\{1, ..., 5\}$ (Fig. 5.20b, 5.20c). Notice that the line is in the boundary of all half planes.

Hence we propose the following:

Conjecture 2. *For the HK dynamics* (5.41) *with topological interaction, for any $d \geq 2$ and $N \geq 4$, the set of equilibria is a stratified manifold with non-separate strata.*



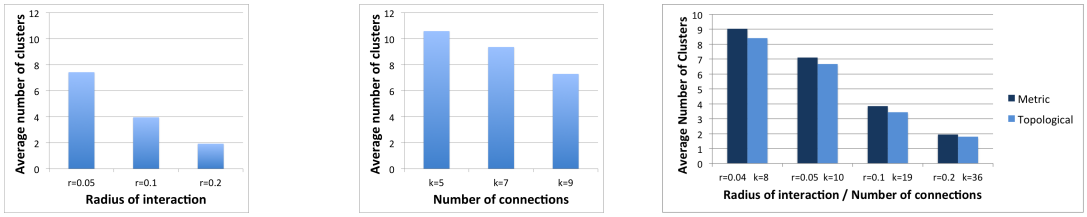
(a) Metric, N=2 (b) Top., k=2, N=5 (c) Top., k=2, N=5

Figure 5.20: Equilibria for the HK system with metric and topological interactions for $d = 1$. Figure (a) shows the equilibrium set in the metric case ($N = 2$), with separate strata. Figure (b) shows the possible configurations for the agents' positions at equilibrium for the topological interaction ($k = 2$, $N = 5$), indicating the number of agents in each cluster and the dimension of the manifold. Figure (c) shows some of the non-separate strata of this equilibrium set.

Numerical results

To compare the different interaction networks, we ran simulations for the well-established one-dimensional HK model, see Figure 5.21. Recent results [6] proposed the idea that topological interactions (with the 5-7 closest neighbors) is an effective way for birds to ensure group cohesion and to escape predators. Figures 5.21a and 5.21b show the average number of clusters of the asymptotic solution of the HK-model (5.41) respectively with metric interaction (5.7) and with topological interaction (5.8), for a group of 100 agents. Notice that consensus is not reached for small radius of interaction ($r \leq 0.2$) or a small number of neighbors ($k < 10$), but instead the group tends to cluster in several subgroups. As expected, the number of clusters decreases as the network connectivity

increases. Figure 5.21c shows that with the same initial number of connections, both interaction networks perform similarly.

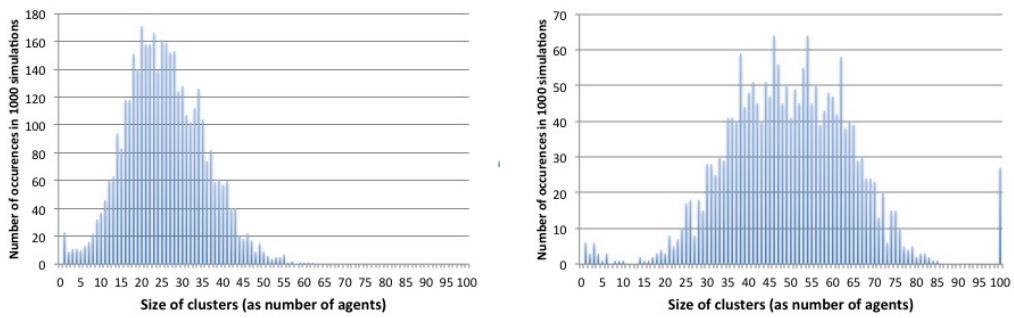


(a) Metric (b) Topological (c) Metric/topological

Figure 5.21: Average number of clusters of the asymptotic solution: (a) for different radii r in the metric configuration, and (b) for different numbers of connections k in the topological configuration. Each average was obtained over 100 simulations, in which 100 agents are initially distributed uniformly in the interval $[0, 1]$. Figure (c) provides a comparison of the two networks, plotting side by side metric and topological configurations with the same initial average number of connections per agent.

In order to illustrate the differently stable equilibrium conformations, we ran simulations with the one-dimensional HK system, plotting the distribution of the asymptotic clusters' sizes (see Figure 5.22). We observed that some conformations are statistically more frequent than others. For instance, in 1000 simulations of the HK dynamics of a group of 100 agents with metric interaction and an interaction radius $r = 0.2$, clusters of 38, 46, 54 and 62 agents are the most frequently obtained (Fig.5.22b). Notice that if $r = 0.2$ and the agents are distributed in the interval $[0, 1]$, there can be at most 4 clusters. We show that in the conditions of the simulations of Fig. 5.22b, in most cases the agents are asymptotically distributed in 2 clusters. Figure 5.23 shows the size distribution of the two biggest clusters (C_1, C_2) over 2000 simulations. The peaks are mostly distributed along the line $C_1 + C_2 = 100$, which means that in most simulations an equilibrium of 2 clusters is reached. Observe that it is less likely to reach an exactly equal distribution of agents between those two clusters than it is to have a slightly unbalanced distribution. The probability of having a very unbalanced distribution decreases with the imbalance.

Long-range connection We verified the effectiveness of long-distance connections in enhancing consensus for social dynamics. For each agent, a distant connection selected uniformly among the other agents was added to each agent's local interactions (see Figure 5.19c). Added to metric interactions, the distant connection almost always lead to consensus. Figure 5.24 shows the improved convergence to consensus when adding an additional distant connection in the HK model. Figure 5.24a shows the evolution of positions with and without an added distant connection. Figure 5.24b shows the evolution of the total number of edges of the network. When distant connections are added, the system asymptotically reaches consensus, and the graph becomes fully connected, i.e. $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ so



(a) $r = 0.1$ (b) $r = 0.2$

Figure 5.22: Distribution of the asymptotic clusters' sizes in 1000 simulations of the one-dimensional HK model with 100 agents and metric interaction. Initially the agents are distributed uniformly in the interval $[0, 1]$. Figure (a) was obtained with an interaction radius $r = 0.1$ and Figure (b) with $r = 0.2$. In the case $r = 0.2$, consensus was reached in 28 simulations. Furthermore, the shape of the distribution suggests that some cluster sizes are more frequent than others.

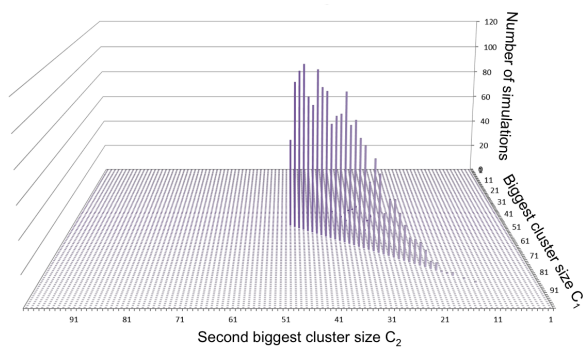
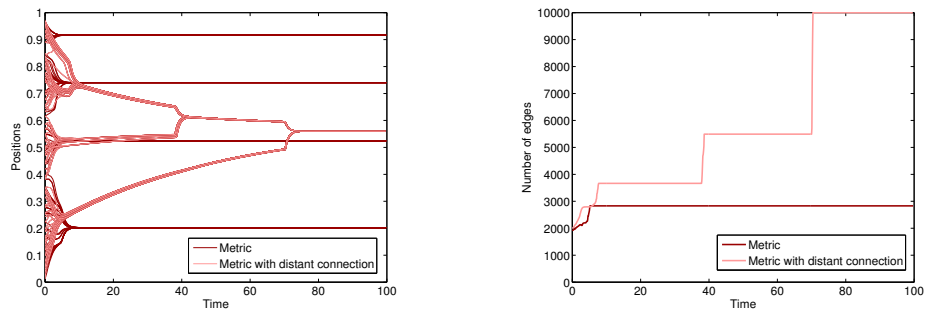


Figure 5.23: Distribution of the two biggest asymptotic clusters' sizes in 2000 simulations of the one-dimensional HK model with 100 agents and metric interaction ($r = 0.2$). Initially the agents are distributed uniformly in the interval $[0, 1]$. The conformation $(C_1, C_2) = (50, 50)$ is obtained in 60 cases, whereas the conformation $(C_1, C_2) = (51, 49)$ is obtained in 100 cases. The most likely conformation is $(C_1, C_2) = (53, 47)$, obtained in 113 cases. There is a low likelihood of having $C_1 - C_2 > 20$.

that $\mathbf{card}(\mathcal{E}) = N^2$.

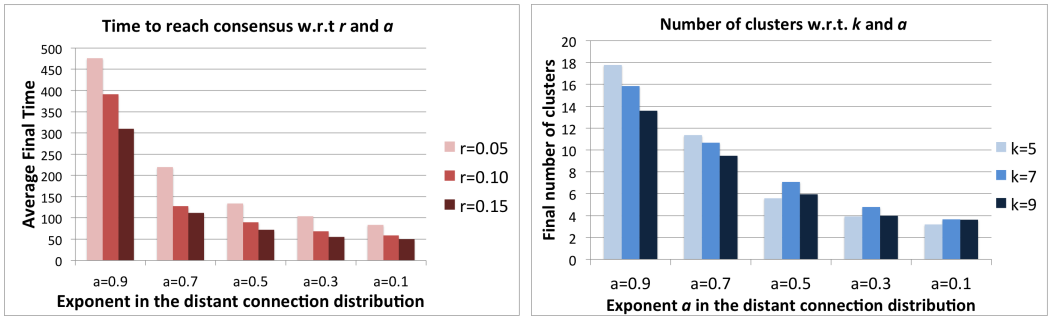
We then studied the effect of the probability with which the distant connection is chosen among all the graph edges. More specifically, we penalize the increase in distance between agents by choosing the distant connection with a probability proportional to ρ^{-a}, where $a \in (0, 1)$ and ρ is the distance between agents. With local metric interaction, adding such a distant connection almost always leads to consensus. With topological interaction, consensus is not always reached but the number of final clusters is significantly reduced. The more biased the choice of distant connection is towards distant neighbors (i.e. the smaller the parameter a), the faster consensus is achieved in the metric case (Fig. 5.25a) or the fewer clusters are obtained in the topological case (Fig. 5.25b).



(a) Evolution of the positions

(b) Evolution of the number of edges

Figure 5.24: Effect of distant connections in convergence to consensus in the HK model with $r = 0.1$. Figure (a) shows the evolution of positions in the metric case with only local interactions or with one added distant connection chosen uniformly (i.e. $a = 0$), resulting respectively in clustering or consensus. Figure (b) shows the evolution of the number of edges.



(a) Time to consensus (metric)

(b) Clustering (topological)

Figure 5.25: Effect of distant connections in convergence to consensus in the HK model. Figure (a) shows the decrease of the time necessary to reach consensus by adding a distant connection (metric case). Since consensus is reached only asymptotically, time to consensus was defined as the time necessary for all agents to be within a sphere of given radius ϵ. Figure (b) shows the decrease of the final number of clusters by adding a distant connection (topological case).

Bibliography

[1] A. Agrachev and Y. Sachkov. *Control Theory from the Geometric Viewpoint*, volume 87. Springer, 2004.

[2] R. Allena, J. J. M. noz, and D. Aubry. Diffusion-reaction model for *Drosophila* embryo development. *Computer Methods in Biomechanics and Biomedical Engineering*, 16(3):235–248, 2013. PMID: 21970322.

[3] A. Aydoğdu, M. Caponigro, S. McQuade, B. Piccoli, N. Pouradier Duteil, F. Rossi, and E. Trélat. Interaction network, state space and control in social dynamics. In N. Bellomo, P. Degond, and E. Tadmor, editors, *Active Particles Volume 1, Theory, Methods, and Applications*. Birkhauser-Springer, 2017.

[4] H.-O. Bae, S.-Y. Ha, Y. Kim, S.-H. Lee, H. Lim, and J. Yoo. A mathematical model for volatility flocking with a regime switching mechanism in a stock market. *Mathematical Models and Methods in Applied Sciences*, 25(07):1299–1335, 2015.

[5] R. E. Baker and P. K. Maini. A mechanism for morphogen-controlled domain growth. *Journal of Mathematical Biology*, 54(5):597–622, 2007.

[6] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1232–1237, 2008.

[7] N. Bellomo, M. A. Herrero, and A. Tosin. On the dynamics of social conflict: looking for the Black Swan. *ArXiv: 1202.4554*, 2012.

[8] N. Bellomo and J. Soler. On the mathematical theory of the dynamics of swarms viewed as complex systems. *Mathematical Models and Methods in Applied Sciences*, 22(supp01):1140006, 2012.

[9] S. Berman, Q. Lindsey, M. Sakar, V. Kumar, and S. Pratt. *Study of group food retrieval by ants as a model for multi-robot collective transport strategies*, volume 6, pages 259–266. MIT Press Journals, 2011.

[10] B. Berret, C. Darlot, F. Jean, T. Pozzo, C. Papaxanthis, and J. P. Gauthier. The inactivation principle: Mathematical solutions minimizing the absolute work and biological implications for the planning of arm movements. *PLOS Computational Biology*, 4(10):1–25, 10 2008.

[11] G. Bliss. The geodesic lines on the anchor ring. *Annals of Mathematics*, 1(4):1–21, 1902.

[12] V. D. Blondel, J. M. Hendrickx, and J. N. Tsitsiklis. Continuous-time average-preserving opinion dynamics with opinion-dependent communications. *SIAM Journal on Control and Optimization*, 48(8):5214–5240, 2010.

[13] A. Bressan and B. Piccoli. *Introduction to the Mathematical Theory of Control*. Springfield, 2004.

[14] A. Bressan and B. Piccoli. *Introduction to the Mathematical Theory of Control*, volume 2 of *AIMS Series on Applied Mathematics*. American Institute of Mathematical Sciences (AIMS), Springfield, MO, 2007.

[15] F. Bullo, J. Cortés, and S. Martínez. Distributed control of robotic networks: a mathematical approach to motion coordination algorithms. *Princeton series in applied mathematics. Princeton University Press, Princeton*, 2009.

[16] S. Camazine, J. Deneubourg, N. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. Self-Organization in Biological Systems. *Princeton University Press*, 2003.

[17] M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and control of the Cucker-Smale model. *Mathematical Control And Related Fields*, 3(4):447–466, 2013.

[18] M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and control of alignment models. *Mathematical Models and Methods in Applied Sciences*, 25(03):521–564, 2015.

[19] M. Caponigro, A. C. Lai, and B. Piccoli. A nonlinear model of opinion formation on the sphere. *Discrete and Continuous Dynamical Systems*, 35(9):4241–4268, 2015.

[20] V. Cavaliere, F. Bernardi, P. Romani, S. Duchi, and G. Gargiulo. Building up the *Drosophila* eggshell: First of all the eggshell genes must be transcribed. *Developmental Dynamics*, 237(8):2061–2072, 2008.

[21] J. Cheeger and D. G. Ebin. *Comparison theorems in Riemannian geometry*, volume 365. AMS Chelsea Publishing, 1975.

[22] L. Chu, H. S. Wiley, and D. A. Lauffenburger. Endocytic relay as a potential means for enhancing ligand transport through cellular tissue matrices: Analysis and possible implications for drug delivery. *Tissue Eng*, 2(1):17–38, 1996.

[23] Y. Chuang, Y. Huang, M. D'Orsogna, and A. Bertozzi. Multi-vehicle flocking: scalability of cooperative control algorithms using pairwise potentials. *IEEE International Conference on Robotics and Automation*, pages 2292–2299, 2007.

[24] I. Couzin and N. Franks. Self-organized lane formation and optimized traffic flow in army ants. *Proc. R. Soc. Lond.*, B 270:139–146, 2002.

[25] I. Couzin, J. Krause, N. Franks, and S. Levin. Effective leadership and decision making in animal groups on the move. *Nature*, 433:513–516, 2005.

[26] I. Couzin, J. Krause, R. James, G. Ruxton, and N. Franks. Collective memory and spatial sorting in animal groups. *J Theor Biol*, 218(1–11), 2002.

[27] E. J. Crampin, E. A. Gaffney, and P. K. Maini. Reaction and diffusion on growing domains: Scenarios for robust pattern formation. *Bulletin of Mathematical Biology*, 61(6):1093–1120, 1999.

[28] E. Cristiani, P. Frasca, and B. Piccoli. Effects of anisotropic interactions on the structure of animal groups. *Journal of mathematical biology*, 62(4):569–588, 2011.

[29] E. Cristiani, B. Piccoli, and A. Tosin. Modeling self-organization in pedestrians and animal groups from macroscopic and microscopic viewpoints. In G. Naldi, L. Pareschi, G. Toscani, and N. Bellomo, editors, *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, Modeling and Simulation in Science, Engineering and Technology. Birkhäuser Boston, 2010.

[30] E. Cristiani, B. Piccoli, and A. Tosin. Multiscale modeling of granular flows with application to crowd dynamics. *Multiscale Model. Simul.*, 9(1):155–182, 2011.

[31] F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Trans. Automat. Control*, 52(5):852–862, 2007.

[32] S. R. X. Dall, L.-A. Giraldeau, O. Olsson, J. M. McNamara, and D. W. Stephens. Information and its use by animals in evolutionary ecology. *Trends in Ecology & Evolution*, 20(4):187–193, 2017/03/30.

[33] M. H. De Groot. Reaching a consensus. *Journal of American Statistical Association*, 69:118 – 121, 1974.

[34] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

[35] J. C. Dittmer. Diskrete nichtlineare modelle der konsensbildung. *Diploma thesis Universität Bremen*, 2000.

[36] F. Dörfler, M. Chertkov, and F. Bullo. Synchronization in complex oscillator networks and smart grids. *Proceedings of the National Academy of Sciences*, 110(6):2005–2010, 2013.

[37] B. Düring, D. Matthes, and G. Toscani. Kinetic equations modelling wealth redistribution: A comparison of approaches. *Phys. Rev. E*, 78:056103, Nov 2008.

[38] M. Erbar. The heat equation on manifolds as a gradient flow in the wasserstein space. *Annales de l'I.H.P. ProbabilitÃľs et statistiques*, 46(1):1–23, 2010.

[39] M. Fornasier, B. Piccoli, and F. Rossi. Mean-field sparse optimal control. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 372(2028), 2014.

[40] J. R. P. French. A formal theory of social power. *Psychological Review*, 63:181 –194, 1956.

[41] W. M. Gelbart, M. Crosby, B. Matthews, W. P. Rindone, J. Chillemi, S. Russo Twombly, D. Emmert, M. Ashburner, R. A. Drysdale, E. Whitfield, G. H. Millburn, A. de Grey, T. Kaufman, K. Matthews, D. Gilbert, V. Strelets, and C. Tolstoshev. Flybase: a *Drosophila* database. the flybase consortium. *Nucleic Acids Research*, 25(1):63–66, 01 1997.

[42] C. Ghiglione, L. Amundadottir, M. Andresdottir, D. Bilder, J. A. Diamonti, S. Noselli, N. Perrimon, and K. L. Carraway III. Mechanism of inhibition of the *Drosophila* and mammalian EGF receptors by the transmembrane protein kekkon 1. *Development*, 130(18):4483–4493, 2003.

[43] I. Giardina. Collective behavior in animal groups: theoretical models and empirical studies. *Human Frontier Science Program Journal*, (205–219), 2008.

[44] C. R. Givens and R. M. Shortt. A class of wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):231–240, 1984.

[45] L. A. Goentoro, G. T. Reeves, C. P. Kowal, L. Martinelli, T. Schüpbach, and S. Y. Shvartsman. Quantifying the gurken morphogen gradient in *Drosophila* oogenesis. *Developmental Cell*, 11(2):263–272, 2002.

[46] J. Gravesen, S. Markvorsen, R. Sinclair, and M. Tanaka. *The cut locus of a torus of revolution.* Technical University of Denmark. Department of Mathematics, 2003.

[47] W. Guo, R. D. Nair, and J.-M. Qiu. A conservative semi-lagrangian discontinuous galerkin scheme on the cubed sphere. *Monthly Weather Review*, 142(1):457–475, 2014.

[48] V. Guttal and I. D. Couzin. Social interactions, information use, and the evolution of collective migration. *Proceedings of the National Academy of Sciences*, 107(37):16172–16177, 2010.

[49] S. Y. Ha, T. Ha, and J. H. Kim. Emergent behavior of a cucker-smale type particle model with nonlinear velocity couplings. *IEEE Transactions on Automatic Control*, 55(7):1679–1683, July 2010.

[50] S.-Y. Ha and E. Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *Kinet. Relat. Models*, 1(3):415–435, 2008.

[51] F. Harary. A criterion for unanimity in french's theory of social power. *Cartwright D (Ed.), Studies in Social Power*, 1959.

[52] L. G. Harrison, S. Wehner, and D. M. Holloway. Complex morphogenesis of surfaces: theory and experiment on coupling of reaction-diffusion patterning to growth. *Faraday Discuss.*, 120:277–293, 2002.

[53] J. Haskovec. Flocking dynamics and mean-field limit in the Cucker–Smale-type model with topological interactions. *Physica D: Nonlinear Phenomena*, 261:42 – 51, 2013.

[54] R. Hegselmann and A. Flache. Understanding complex social dynamics – a plea for cellular automata based modelling. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998.

[55] R. Hegselmann, U. Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[56] R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann. Probab.*, 3(4):643–663, 08 1975.

[57] D. Horstmann. From 1970 until present: The Keller-Segel model in chemotaxis and its consequences. I. *Jahresber. Dtsch. Math.-Ver.*, 105(3):103–165, 2003.

[58] D. Horstmann. From 1970 until present: the Keller-Segel model in chemotaxis and its consequences. II. *Jahresber. Dtsch. Math.-Ver.*, 106:51–69, 2004.

[59] A. Huth and C. Wissel. The simulation of the movement of fish schools. *Journal of Theoretical Biology*, 156:365–385, 1992.

[60] A. Jadbabaie, J. Lin, and A. S. Morse. Correction to: "Coordination of groups of mobile autonomous agents using nearest neighbor rules" [IEEE Trans. Automat. Control **48** (2003), no. 6, 988–1001; MR 1986266]. *IEEE Trans. Automat. Control*, 48(9):1675, 2003.

[61] E. F. Keller and L. A. Segel. Initiation of slime mold aggregation viewed as an instability. *J. Theor. Biol.*, 26(3):399–415, 1970.

[62] J. Kelley. *General topology.* Springer-Verlag, 1975.

[63] J. M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 08 2000.

[64] S. Kondo and R. Asai. A reaction-diffusion wave on the skin of the marine angelfish pomacanthus. *Nature*, 376(6543):765–768, 08 1995.

[65] J. Krause and G. Ruxton. Living in groups. *Oxford series in ecology and evolution. Oxford University Press, New York*, 2002.

[66] U. Krause. Soziale dynamiken mit vielen interakteuren, eine problemskizze. *Krause U and Stöckler M (Eds.) Modellierung und Simulation von Dynamiken mit vielen interagierenden Akteuren, Universität Bremen*, pages 37 – 51, 1997.

[67] U. Krause. A discrete nonlinear and non—autonomous model of consensus formation. *Elaydi S, Ladas G, Popenda J and Rakowski J (Eds.), Communications in Difference Equations, Amsterdam: Gordon and Breach Publ.*, pages 227 – 236, 2000.

[68] Y. Kuramoto. Cooperative dynamics of oscillator community a study based on lattice of rings. *Progress of Theoretical Physics Supplement*, 79:223–240, 1984.

[69] B. Laroche, P. Martin, and P. Rouchon. Motion planning for the heat equation. *International Journal of Robust and Nonlinear Control*, 10(8):629–643, 2000.

[70] J.-M. Lasry and P.-L. Lions. Mean field games. *Jpn. J. Math. (3)*, 2(1):229–260, 2007.

[71] J. Lefèvre and J.-F. Mangin. A reaction-diffusion model of human brain development. *PLOS Computational Biology*, 6(4):1–10, 04 2010.

[72] K. Lehrer. Social consensus and rational agnoiology. *Synthese*, 31:141 – 160, 1975.

[73] N. Leonard. Multi-agent system dynamics: Bifurcation and behavior of animal groups. *Plenary paper IFAC Symposium on Nonlinear Control Systems, Toulouse, France.*, 2013.

[74] N. Leonard and E. Fiorelli. Virtual leaders, artificial potentials and coordinated control of groups. *Proc. 40th IEEE Conf. Decision Contr.*, pages 2968–2973, 2001.

[75] N. E. Leonard. Multi-agent system dynamics: Bifurcation and behavior of animal groups. *Annual Reviews in Control*, 38(2):171 – 183, 2014.

[76] P. K. Maini, T. E. Woolley, R. E. Baker, E. A. Gaffney, and S. S. Lee. Turing's model for biological pattern formation and the robustness problem. *Interface Focus*, 2(4):487–496, 08 2012.

[77] A. Matikas, D. Mistriotis, V. Georgoulias, and A. Kotsakis. Current and future approaches in the management of non-small-cell lung cancer patients with resistance to EGFR TKIs. *Clinical Lung Cancer*, 16(4):252 – 261, 2015.

[78] T. Miura, K. Shiota, G. Morriss-Kay, and P. K. Maini. Mixed-mode pattern in doublefoot mutant mouse limb—turing reaction–diffusion model on a growing domain during limb development. *Journal of Theoretical Biology*, 240(4):562 – 573, 2006.

[79] S. A. Molchanov. Diffusion processes and riemannian geometry. *Russian Mathematical Surveys*, 30(1):1, 1975.

[80] C. Moore. Braids in classical dynamics. *Physical Review Letters*, 70:3675–3679, June 1993.

[81] C. Moore and M. Nauenberg. New periodic orbits for the n-body problem. *ASME. J. Comput. Nonlinear Dynam.*, 4(1):307–311, 2006.

[82] S. Motsch and E. Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.

[83] R. D. Nair, S. J. Thomas, and R. D. Loft. A discontinuous galerkin transport scheme on the cubed sphere. *Monthly Weather Review*, 133(4):814–828, 2005.

[84] F. Neuman-Silberberg and T. Schüpbach. Dorsoventral axis formation in *Drosophila* depends on the correct dosage of the gene gurken. *Development*, 120(9):2457–2463, 1994.

[85] F. S. Neuman-Silberberg and T. Sch˙ The *Drosophila* dorsoventral patterning gene gurken produces a dorsally localized rna and encodes a TGFα-like protein. *Cell*, 75(1):165 – 174, 1993.

[86] M. G. Niepielko and N. Yakoby. Evolutionary changes in TGFα distribution underlie morphological diversity in eggshells from *Drosophila* species. *Development*, 141(24):4710–4715, 2014.

[87] H. Niwa. Self-organizing dynamic model of fish schooling. *J. Theor. Biol.*, 171:123–136, 1994.

[88] J. Parrish and L. Edelstein-Keshet. Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science*, 294:99–101, 1999.

[89] J. Parrish, S. Viscido, and D. Gruenbaum. Self-organized fish schools: An examination of emergent properties. *Biol. Bull.*, 202:296–305, 2002.

[90] C. S. Patlak. Random walk with persistence and external bias. *Bull. Math. Biophys.*, 15:311–338, 1953.

[91] L. Perea, G. Gómez, and P. Elosegui. Extension of the Cucker-Smale control law to space flight formations. *AIAA Journal of Guidance, Control, and Dynamics*, 32:527–537, 2009.

[92] F. Peri, C. Bökel, and S. Roth. Local Gurken signaling and dynamic MAPK activation during *Drosophila* oogenesis. *Mechanisms of Development*, 81(1-2):75 – 88, 1999.

[93] B. Perthame. *Transport Equations in Biology.* Basel: Birkhäuser, 2007.

[94] B. Piccoli, N. Pouradier Duteil, and B. Scharf. Optimal control of a collective migration model. *Mathematical Models and Methods in Applied Sciences*, 26(02):383–417, 2016.

[95] B. Piccoli and F. Rossi. Transport equation with nonlocal velocity in wasserstein spaces: Convergence of numerical schemes. *Acta Applicandae Mathematicae*, 124(1):73–105, 2013.

[96] B. Piccoli and F. Rossi. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

[97] I. Pinilla-Macua, S. C. Watkins, and A. Sorkin. Endocytosis separates EGF receptors from endogenous fluorescently labeled hras and diminishes receptor signaling to map kinases in endosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(8):2122–2127, 02 2016.

[98] R. Plaza, F. Sánchez-Garduño, P. Padilla, R. Barrio, and P. Maini. The effect of growth and curvature on pattern formation. *Journal of Dynamics and Differential Equations*, 16(4):1093–1121, 2004.

[99] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes.* Interscience Publishers John Wiley & Sons, Inc. New York-London, 1962.

[100] N. Pouradier Duteil, F. Rossi, U. Boscain, and B. Piccoli. Developmental partial differential equations. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 3181–3186, Dec 2015.

[101] M. Přibyl, C. B. Muratov, and S. Y. Shvartsman. Discrete models of autocrine cell communication in epithelial layers. *Biophysical Journal*, 84(6):3624–3635, 06 2003.

[102] M. Přibyl, C. B. Muratov, and S. Y. Shvartsman. Long-range signal transmission in autocrine relays. *Biophysical Journal*, 84(2):883–896, 02 2003.

[103] W. Romey. Individual differences make a difference in the trajectories of simulated schools of fish. *Ecol. Model.*, 92:65–77, 1996.

[104] F. Rossi, N. Pouradier Duteil, N. Yakoby, and B. Piccoli. Control of reaction-diffusion equations on time-evolving manifolds. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1614–1619, Dec 2016.

[105] J. S. Rush, L. M. Quinalty, L. Engelman, D. M. Sherry, and B. P. Ceresa. Endosomal accumulation of the activated epidermal growth factor receptor (EGFR) induces apoptosis. *Journal of Biological Chemistry*, 287(1):712–722, 2012.

[106] A. Sarlette, S. Bonnabel, and R. Sepulchre. Coordinated motion design on lie groups. *Automatic Control, IEEE Transactions on*, 55(5):1047–1058, May 2010.

[107] A. Sarlette and R. Sepulchre. Consensus optimization on manifolds. *SIAM Journal on Control and Optimization*, 48(1):56–76, 2009.

[108] L. Scardovi, A. Sarlette, and R. Sepulchre. Synchronization and balancing on the N-torus. *Systems & Control Letters*, 56(5):335 – 341, 2007.

[109] T. Schüpbach. Germ line and soma cooperate during oogenesis to establish the dorsoventral pattern of egg shell and embryo in *Drosophila* melanogaster. *Cell*, 49(5):699–707, 1987.

[110] S. Seirin Lee, E. A. Gaffney, and R. E. Baker. The dynamics of turing patterns for morphogen-regulated growing domains with cellular response delays. *Bulletin of Mathematical Biology*, 73(11):2527–2551, 2011.

[111] R. Sepulchre, D. Paley, N. E. Leonard, et al. Stabilization of planar collective motion: All-to-all communication. *Automatic Control, IEEE Transactions on*, 52(5):811–824, 2007.

[112] R. Sepulchre, D. Paley, N. E. Leonard, et al. Stabilization of planar collective motion with limited communication. *Automatic Control, IEEE Transactions on*, 53(3):706–719, 2008.

[113] J. Shatah and M. Struwe. The cauchy problem for wave maps. *International Mathematics Research Notices*, 2002(11):555, 2002.

[114] B.-Z. Shilo. Regulating the dynamics of EGF receptor signaling in space and time. *Development*, 132(18):4017–4027, 2005.

[115] S. Sigismund, E. Argenzio, D. Tosoni, E. Cavallaro, S. Polo, and P. P. D. Fiore. Clathrin-mediated internalization is essential for sustained EGFR signaling but dispensable for degradation. *Developmental Cell*, 15(2):209 – 219, 2008.

[116] P. Sobkowicz. Modelling opinion formation with physics tools: Call for closer link with reality. *Journal of Artificial Societies and Social Simulation*, 12(1):11, 2009.

[117] E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems, Second Edition*. Springer, New York, 1998.

[118] A. Spradling. *Developmental genetics of oogenesis*. M. Bate, A.M. Arias (Eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1993.

[119] S. H. Strogatz. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1–4):1 – 20, 2000.

[120] M. Struwe. *Variational Methods. Applications to Nonlinear Partial Differential Equations and Hamiltonian Systems*.

[121] K. Sugawara and M. Sano. Cooperative acceleration of task performance: Foraging behavior of interacting multi-robots system. *Physica D*, 100:343–354, 1997.

[122] D. Sumpter. The principles of collective animal behaviour. *Philosophical Transaction of the Royal Society B*, 361:5–22, 2006.

[123] N. Taleb. *The Black Swan*. Penguin, 2010.

[124] J. Toner and Y. Tu. Long-range order in a two-dimensional dynamical xy model: How birds fly together. *Phys. Rev. Lett.*, 75:4326–4329, 1995.

[125] A. M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 237(641):37–72, 1952.

[126] C. Varea, J. L. Aragón, and R. A. Barrio. Confined turing patterns in growing systems. *Phys. Rev. E*, 56:1250–1253, Jul 1997.

[127] C. Varea, J. L. Aragón, and R. A. Barrio. Turing patterns on a sphere. *Phys. Rev. E*, 60:4588–4592, Oct 1999.

[128] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.*, 75:1226–1229, Aug 1995.

[129] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, 2008.

[130] D. Vilone, J. Ramasco, A. Sánchez, and M. S. Miguel. Social and strategic imitation: the way to consensus. *Scientific Reports*, 2, 09 2012.

[131] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 06 1998.

[132] H. Whitney. On singularities of mappings of Euclidean spaces. I. mappings of the plane into the plane. *Annals of Mathematics*, 62(3):374–410, 1955.

[133] J. J. Zartman, L. S. Cheung, M. Niepielko, C. Bonini, B. Haley, N. Yakoby, and S. Y. Shvartsman. Pattern formation by a moving morphogen source. *Physical biology*, 8(4):045003–045003, 08 2011.