Congestion in many-particle systems with volume exclusion constraints: algorithms and applications to modelling in biology

a thesis presented for the degree of

Doctor of Philosophy of Imperial College London

and the

Diploma of Imperial College

by

Marina Amado Ferreira

Department of Mathematics

Imperial College

180 Queen's Gate, London SW7 2BZ

June 2018

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: Marina Ferreira

Copyright

Congestion in many-particle systems with volume exclusion constraints: algorithms and applications to modelling in biology

Abstract

Many-particle systems with congestion are widely found in biology, for example, in cell tissues or herds. Mathematical modelling constitutes an important tool in their study. In contrast to common approaches, we propose two new modelling frameworks that rely on the exact treatment of the contacts between particles: a particle-based and a continuum framework. Both frameworks are based on the same behavioural rules, namely 1) two particles cannot overlap with each other and 2) the particles seek a minimum of a given confining potential at all times. The dynamics is driven by the evolution of the potential and changes in particle characteristics, such as size.

In the first part, the static equilibria of the particle-based model are obtained as solutions to a minimization problem. This leads to non-convex optimization under volume exclusion constraints. Classical tools are either not applicable or not efficient. We develop and study a new and efficient minimization algorithm to approximate a solution.

The second part concerns the time-evolution of the particle-based framework. We develop new time-stepping schemes involving the resolution of a minimization problem at each time-step, which is tackled with the minimization algorithm developed in the first part. The study of these schemes is performed in the case of a system of hard-spheres undergoing ballistic aggregation on a torus and it succeeds to simulate up to one million particles. These new tools are applied to the study of the mechanics of a cell tissue, which has allowed to validate them in practice.

In the third part, we develop a continuum modelling framework describing the evolution of particle density. Our approach differs from previous ones by relying on different modelling assumptions that are more appropriate to biological systems. We show that this novel approach leads to a free-boundary problem and we characterize the dynamics of the boundary.

Keywords: Volume exclusion constraints, confining potential, packing problems, non-convex optimization, time-stepping scheme, cell tissue mechanics, free-boundary problem.

To my parents, Maria and Martinho.

To the memory of my grandfather, Serafim.

Acknowledgments

First I would like to express my deep gratitude to my supervisors Pierre Degond and Sara Merino-Aceituno for supporting, guiding and inspiring me all the way during my PhD. Pierre has patiently and persistently guided me through a challenging beginning. All his efforts turned out to be successful and I can't say how much I am thankful for this. Along the PhD he had wisely driven me towards becoming an independent researcher, while being always available to answer my questions and to discuss and support my own ideas. Pierre has introduced me to the fascinating world of particle systems and to the challenges of modelling, simulating and analysing them. It has been a honour and a great enjoyment to work with Pierre. His wide knowledge besides mathematics has enriched my view of the mathematical problems arising in biology and contributed to the success of the collaboration with biologists. Sara is a source of positive energy. She always knew how to encourage and guide me through the less good moments. I want to thank her for the very valuable advices on how to write scientific papers and how to manage my time, as well as, on scientific guidance, especially during the writing up stage of my PhD.

I would like to thank Sebastien Motsch (ASU) for welcoming me in Arizona and for all the fruitful discussions during my stay there. Also, I want to thank Eric Theveneau (CNRS, Toulouse) for introducing me to his lab and their research and for being patient enough to answer to my naive questions about cell tissues. This effort has certainly been rewarded by the exciting achievements we got together and the future perspectives we envision. Thanks also to Cyril Andrieu, Fernanda Bajanca, Anne Bibonne, Angie Molina and especially to Nadège Gouignard for being like a family to me during my stay in Toulouse. I would like to thank to Giacomo Dimarco (Ferrara) for the insightful discussions on the numerical methods for ballistic aggregation and to Mattias Zanella for the time in Ferrara and for showing around the town.

I am extremely grateful to Robert Nürnberg and Eric Keaveny for the very useful and valuable comments and advices during the assessment points throughout the PhD. A big thank you to Diane Peurichard (UPMC) for introducing me to the computer language *Fortran* and for being always available to help in any problem I had. Thank you to my examiners Bertrand Maury (UPS) and Demetrios Papageorgiou (ICL) for carefully reviewing this dissertation.

I would like to thank to Sophie Hecht, Ewelina Zatorska, Pedro Aceves Sanchez, Sara Merino-Aceituno and Pierre Degond for all advices, motivation and good moments at Imperial during the numerous lunches at the market and tea and chocolate at Pierre's office. I will especially miss this last part! I would like to thank to José Carrillo for

inviting me to present my work at Imperial and for introducing me to Pierre. A big thank you to my officemates and colleagues from Imperial Franca Hoffmann, Anna Ananova, Madeleine Whybrow, Thomas Johnson, Sara Algeri, Urbain Vaes, Xue Lu (Lucy), Marco Francischello, Luca Mingarelli, Silvia Metelli, Onur Teymur, Maximilian Engel, Nikolas Nuesken, Martin Weidner, Cezary Olszowiec, Samuel Brzezicki, Sophie Hecht, Adam Gosztolai, Susana Gomes and Radu Cimpeanu for sharing and supporting this experience of being a PhD student. Thanks also for the conversations, trips, beers, football, cinema, concerts, museums and so on.

Thanks to the SIAM Chapter committee Franca Hoffmann, Adam Gosztolai, Alex Amos, Hanne Hoitzing, Michael Cornish, Juvid Aryaman and Arman Khaledian for sharing both the hard work and the fun of organizing scientific and social activities in the department. Thanks to the student reps Chiara Taranto, Matt Price-Williams and Arman Khaledian for their will to contribute to a stronger sense of community within the department. Also a big thanks to Anderson Santos for supporting and helping in all these activities. Thanks to all staff in general of the Maths Department for all the technical and logistic support.

Thanks to Susana Gomes for having welcoming me in Imperial, for introducing me to José and Pierre and for making the beginning of my stay in London smooth. Thanks also for all the valuable advices regarding conferences or other research-related topics and for all amazing hikes and dinners. Thanks to my Portuguese friends Rita Ferreira, Rui Sequeira, Verónica Silva, Ana Pereira, Eliana Silveira and Dora Pestana for all the incredible moments we have been through since undergrads and for still keeping in touch as if the years hadn't passed. Thanks to Flávio Cordeiro for the important support and advices during the first stage of the PhD.

I want to thank Philip Maini (Oxford) for the inspiring talks at the AARMS Summer School in St. John, Newfoundland, Canada, which drove me into the PhD. Thanks also for all advices and encouragement. A big thank you to Maria Paula de Oliveira and Ana Cristina Rosa, my master thesis supervisors at the University of Coimbra, for introducing me to partial-differential equations and stochastic differential equation arising in biology and for guiding me during my first steps doing research. Also to Leal Duarte, João Queiró, Paulo Oliveira, Maria Celeste Gouveia and Alexander Kovacec from Coimbra who showed me the beauty of mathematics and provided me with a solid basis which allowed me to proceed to a research career.

A special thanks to Filipe Carvalho for sharing all life adventures with me since high school, for encouraging me to pursue my goals and at the same time being able to cope with my absence while I am doing so. Without his incentive and support this PhD would not have been possible. Thanks also for the super interesting discussions on biology and ecology that roused my earliest interest for mathematical problems in biology.

Many thanks to my family for always encouraging and cheering me up. To my parents

and dear siblings for their support, strength and faith. To my grandparents and uncles for having to deal with my absence and for their constant concern and care.

Muito obrigada à minha família por me incentivarem e animarem sempre. Aos meus pais e queridos irmãos pelo apoio, força e fé. Aos meu avós e tios por terem de lidar com a minha ausência e pela sua preocupação e carinho constantes.

Financial support

I would like to acknowledge financial support from the Department of Mathematics of Imperial College through a Roth PhD Studentship. Additional funding has been provided to support my participation in conferences and research visits. In particular, I acknowledge support from Imperial College through a Doris Chen Mobility award to visit Sebastien Motsch in Arizona, The Company of Biologists to visit Theveneau's lab in Toulouse and Institute of Mathematics and its Applications to attend a CIMPA School in Cuba. I would like to thank Eric Theveneau, Sebastien Motsch and Giacomo Dimarco for supporting my visits to their research groups.

List of figures

List of tables

Contents

0 Introduction

0.1 Overview

A many-particle system consists of a collection of particles, individuals or agents that interact with each other and with the surrounding environment. A few examples include molecules, granular material, such as dust or sand, cell tissues, herds or crowds of people. In some instances, if the density of the particles is very high, the system becomes *congested* or *densely packed* and contacts or near-contacts between particles become non-negligible phenomena. In Section 0.2 we present a few examples of congested systems. The study of congested systems requires the use of mathematical models that take into account the volume of the particles, through for example *volume exclusion* or *non-overlapping* constraints.

A many-particle system, or simply, a particle system exhibits *emergent behaviour* when macroscopic features appear as a result of the interaction between particles, without being directly encoded in them. To account for this behaviour, one needs to develop micro- and macroscopic models of the same system that are consistent with each other. This motivates the problem of finding a rigorous connection between these models, which constitutes a big challenge in mathematics [76]. In this thesis we derive a microscopic (Chapters 1-4) and a macroscopic (Chapter 5) modelling framework for congested systems that share the same *rules* or *principles*. This work intends to be a first step towards a rigorous connection between the two scales for a particular case of congested systems that we will detail in this thesis. The microscopic model, which typically corresponds to a *discrete* or *particle-based* model, describes the evolution of a finite number of particles, while the macroscopic model, corresponding to *continuum* or *density based* model, describes the evolution of a continuous particle density.

The rules underlying the dynamics of certain biological systems, such as cell tissues,

may be very complex. In order to build appropriate models for such systems, rather than using physical first principles, we will formulate heuristic rules based on simple hypotheses about the underlying particle behaviour. The two main rules that we will consider here are that 1) particles cannot overlap and 2) they seek at all times a state of minimal energy. This translates into the minimization of an energy, which is given by a confining potential, subject to non-overlapping constraints.

In the following we describe the main challenges arising in the microscopic and macroscopic modelling of congested systems and how we approach them in this thesis. Additionally, we motivate the application of this framework to the study of a packed cell tissue.

In models at the microscopic scale, space may be regarded as discrete or continuous. The first case leads to on-lattice models where the particles occupy one or several units of a grid at all times. In such models the volume exclusion constraints are easily ensured, nevertheless, these models are not appropriate to address congestion due to the effects introduced by the grid as it is reported in [77, 126, 150]. In our approach we rather consider an off-lattice model where the particles can occupy any position in the Euclidean space. For this class of models, the exclusion volume constraints may be approximated by a soft repulsion potential. Despite being computationally less expensive, this approximation may become less and less accurate as the compression forces generated by congestion increase. To avoid this situation we use hard-constraints that are enforced at all times. In our model, the particles are represented by hard-spheres that move over time in a given domain and interact with each other according to certain collision rules. Such interactions lead to discontinuities in the velocity (non-smooth dynamics), which makes the dynamics of the system very complex, especially when the number of particles is large. In order to get some insight about its evolution, we resort to computer simulations as explained next.

There are two types of schemes to simulate non-smooth systems [20, 111]: event-driven (ED) and time-stepping (TS) (see Section 0.3.2.2 for a description of the main algorithms presented in the literature). In the first case, the following two steps are successively performed: 1) smooth evolution of the system until the next non-smooth event (for example, a contact event) and 2) determination of the velocities after the event. Despite being more accurate, this class of schemes becomes computationally very expensive as the rate of events increases, which is likely to happen in a congested system with many particles. In this situation, a time-stepping scheme may therefore be more appropriate, as we observe in the study of ballistic aggregation presented in Chapter 3. In a TS scheme, the time is discretized in regular intervals and, at each time-step, the particles evolve according to the smooth dynamics, which may lead to overlapping with neighbouring particles. Since this configuration is non-admissible, a nearby admissible configuration is then obtained as a minimizer of a given potential energy subject to non-overlapping constraints. Such a configuration represents a mechanical equilibrium of the system. In Chapter 3 we propose a new TS method which is similar to the class of

methods proposed by Schatzmann and Paoli [139, 140] in the context of impact problems with unilateral contacts and Coulomb's friction (see Section 0.3.2.2 for a brief discussion on the TS methods proposed in the literature). We show in the context of ballistic aggregation that the TS method is performing better than the ED method when the volume fraction of the particles is large. Indeed the simulation with the TS scheme takes a few hours while the ED scheme takes a few days.

The main difficulty of the TS scheme lies on the minimization problem with non-overlapping constraints. Indeed, the non-overlapping constraints typically lead to non-convex optimization (see Section 0.3.1.2 for an elaboration of this point). Unfortunately there are no universal tools available and standard methods turn out to be computationally expensive when the number of particles is large (see Section 0.3.2.1 for a brief overview on the available algorithms for non-convex optimization). We then propose a new algorithm to tackle this problem: the damped Arrow-Hurwicz algorithm (DAHA). In Section 0.3.2.1 we provide the necessary conditions for optimality that are on the basis of this algorithm. Part I of this thesis is dedicated to the development and analysis of the DAHA. Specifically, in Chapter 1 we present the derivation of this method and show that it performs better than existing methods in terms of computational speed in the case of a packing problem with a large number of particles. In Chapter 2 we present an analytical study on the convergence of the method in one spatial dimension as well as a parametric investigation of the convergence speed in any spacial dimension.

The model and algorithms developed for the microscopic setting are applied to the study of a biological system: an epithelial tissue. An epithelial tissue is constituted of closely packed cells which adhere with their neighbours. This system represents an example of a congested system and therefore we may use our framework to model it. A more elaborated justification for this choice is presented in Section 0.3.2.2, as well as a brief overview on existing models. The biological motivation is to try to identify the main mechanisms underlying the dynamics and shape of the tissue. In particular, in the case of the healthy tissue, how the tissue shape remains stable despite of all the cell movements and deformations and, in the case of a cancer tissue, how is the tissue shape disrupted when defects in individual cells are introduced. In this thesis we primarily focus on the healthy tissue. The results obtained in close collaboration with a team of biologists (Eric Theveneau's team, Toulouse) are presented in Chapter 4. The model consists of a system of spheres with a hard inner-core and a soft outer-core, that represent the cell nucleus. The spheres are connected to each other through springs, representing the cytoplasm. *In silico* and *in vivo* experiments have been run in parallel. As a result, we have been able to verify some hypotheses regarding the mechanics of the system. Specifically, the model suggests that the soft-core of the nucleus and the presence of noise are essential features, without which the system becomes too rigid and is not able to reproduce the real dynamics observed *in vivo*.

At the macroscopic level, rather than looking at the behaviour of each particle, we look

at the evolution of the density of particles in space. We assume heuristic rules that we have highlighted before namely 1) particles cannot overlap and 2) they seek at all times a state of minimal energy. This equilibrium state is a minimizer of a confining potential energy subject to maximal density constraints. The constraints depend on the average particle volume, which may grow as a consequence of the particle swelling. In Section 5.3 we analyse the constrained minimization problem and show that its solution leads to a free-boundary problem where the dynamics of the boundary is driven by the evolution of the confining potential and of the average particle volume. This work is presented in Chapter 5. Similar models have been proposed in the literature (see Section 0.3.3.2 for a brief overview), such as Hele-Shaw type models. They differ from our approach by relying on different physical assumptions, which leads to different predictions on the behaviour of the material. Our approach constitutes therefore a new modelling framework for congested systems and provides an attractive basis for the development of complex models for swelling materials, such as tumours.

The present Chapter 0 comprehends two parts. The first part is constituted by Sections 0.2 and 0.3 and contains an overview of the main topics discussed in this thesis and of the tools needed. Specifically, in Section 0.2 we present a general description of congested systems and examples and in Section 0.3 we provide an overview on mathematical models and algorithms for congested systems. Some background theory of optimization that is essential in our modelling framework is also presented in Sections 0.3.2.1 and 0.3.3.1 . The second part contains a summary of the main results and contributions to this thesis (Section 0.4) and a schematic representation of this thesis (Section 0.5).

0.2 Congested systems with contact events

In Section 0.2.1 we provide a description of congested systems. In particular we describe the different types of local interactions that distinguishes different types of systems. We then present the systems that are discussed in this thesis (see Section 0.2.2) and other systems that could be studied using our mathematical framework and computational tools (see Section 0.2.3).

0.2.1 Description

A *congested system* is a many-particle system where the particles are densely distributed in space and contacts or near-contacts between particles have a non-negligible effect on the static state or evolution of the system. Congested systems may differ on the properties of the particles (for example shape, growth, shrinkage, stiffness, preferred direction or random behaviour) and on the type of interactions they are subjected to. These interactions may come from the other particles of the system (for example, collisions, repulsion forces or attraction forces) or from the environment (for example, chemical signals, population

behaviour or space constraints).

Congested systems can be found for instance in vehicle traffic, crowd evacuation, cell tissues or granular matter. In this thesis we focus on *congested systems with contacts*. In the following paragraphs we present the systems treated in this thesis (Section 0.2.2) as well as other systems that could be studied using our modelling and computational tools (Section 0.2.3).

0.2.2 Examples treated in this thesis

0.2.2.1 Sphere packing

In Chapters 1 and 2, we address the question of given an arbitrarily finite sampling of spheres (possibly overlapping), how to get a (non-overlapping) packed configuration that is close to the given one. We address this question through the development and implementation of algorithms for non-convex optimization. Figure 1 shows a possible solution to this problem. Note that we are not necessarily interested in the global optimum, as in many natural systems individuals have only local information and therefore they only look for a local optimizer.

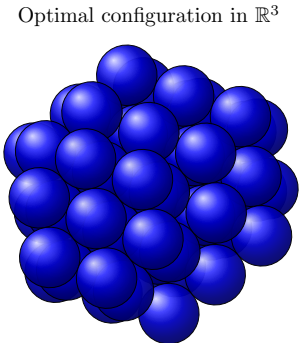


Figure 1: Example of a sphere packing in three spatial dimensions. This configuration has been generated with the damped Arrow-Hurwicz algorithm detailed in Chapter 1.

Optimization problems associated to sphere packing in a container have been reviewed in [89]. Packings of generic shaped particles have also been widely investigated in the context of granular media [117], glasses [181], liquids [85], biology (bacterial colony growth [107]) and in social sciences (maximum safe packing of living agents in crowds or containers [39]). In geometry, the old Kepler problem of finding the densest packing of spheres in the Euclidean space has only been solved in 2005 in dimension 3 [82, 178] and more recently, in dimensions 8 and 24 [36].

0.2.2.2 Ballistic aggregation

Ballistic aggregation consists of a system of rigid particles moving in straight trajectories until they collide, stick and form growing aggregates [56, 175]. The inelastic collision of two

aggregates conserves the total mass and momentum, while kinetic energy is dissipated. Note that given the initial positions and velocities, this dynamics is deterministic. In Chapter 3 we consider a model for ballistic aggregation in two spatial dimensions with periodic boundary conditions (see Figure 2). We use this model to explore and compare several algorithms in the case of a large number of particles (up to one million).

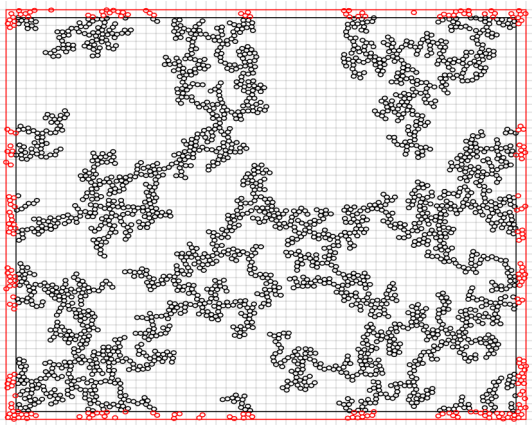


Figure 2: Example of a final cluster obtained from ballistic aggregation dynamics on a domain with periodic boundary conditions. This configuration has been generated with the time-stepping algorithm with non-smooth constraints and non-rigid clusters, NS-NR, detailed in Chapter 3.

Models for ballistic aggregation have been used in the study of evaporated target foils [169], silica-sol gel films [78], a granular gas [142], cosmic dust [120] and protoplanetary disks [136]. Other types of aggregation include diffusion-limited aggregation [13, 63, 156, 175], coalescence [6, 67], among others [121, 125].

0.2.2.3 Epithelial tissue mechanics

Epithelial tissues constitute one of the four types of animal tissue together with connective tissue (adipose tissue, bones, blood, cartilage, etc), muscle tissue and nervous tissue. They are constituted of closely packed cells that are adjacent to a basement membrane (see Figure 3). These tissues are found in the whole body and constitute the surface of organs and blood vessels [80]. For this reason, they have been studied in the context of embryonic development [83] to understand how organs acquire their shape and how it remains stable, and in the context of cancer, to understand how the tissue shape is disrupted [172]. Indeed, many tumours have their origin in epithelial tissues.

Such questions have been explored in the literature with the aid of mathematical models. In particular, since epithelial tissues are packed, the underlying mechanics, such as the contact forces between neighbouring cells, usually play a central role in these studies [8]. Models of epithelial tissue range from individual based [14] to continuum models [92, 105, 176]. Individual based models describe the tissue at the cell level and study the local behaviour, for example, how the variation of spatial constraints influence the cell cycle [170] or how contact inhibition of locomotion originates forces in the tissue

[182]. See Section 0.3.2.2 for a brief overview of particle based models and see [138] for a comparison between them. Continuum models instead describe the system at the tissue level (cell density) and focus on global properties of the tissue, such as the curvature, resistance to deformation [8], contraction-elongation and tissue shear flow [92], to name a few. Both types of models may be coupled with models for the concentration of molecules [177] or subcellular structures, such as microtubules and actin filaments, leading to multiscale models [33, 42]. The mathematical connection between individual based and continuum models has been developed in [19] for the case of two interacting populations of cells with flux-limited chemotaxis.

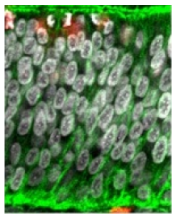


Figure 3: Example of an epithelial tissue obtained from the chick embryo by Theveneau's team in Toulouse (see Chapter 4). We see a high density of cell nuclei.

In Chapter 4 we build a particle based model for the mechanics of an epithelial tissue (see Figure 3) in order to study the effect of congestion on the morphology and dynamics of the tissue.In Chapter 5 we build a continuum model for the cell density which includes cell growth and movement under a confining potential and maximal density (packing) constraints.

0.2.2.4 Swelling materials

In swelling or drying materials the volume of the material increases or decreases, respectively, due to the absorption or loss of some substance (typically water). These materials have been studied in many different contexts for example in biology (mitotic cells or tumours [11, 183]), chemistry or material science (swelling gels [114]), geosciences (drying of wetting soil [68]) and food (potato starch swelling [93]).

Mathematical models of swelling material have been proposed in the context of tumour growth, namely in the study of malignant melanoma [11] or in the study of the relation between differential cell growth and the mechanical stress experienced by cells in a tumour [97]. Mathematical models have also been used to model the rate and degree of swelling of gels in the context of drug delivery and tissue repair [73].

In Chapter 5 we develop a continuum model for incompressible swelling materials based on simple rules on the interactions between the constituent particles. The main rule is that the particles cannot overlap. The model describes the evolution of the number density of particles that results from the swelling or shrinkage of the particles and from the evolution of a given confining potential.

0.2.3 Other examples

In this Section we briefly present other congested systems that could be studied using the tools developed in this thesis, even though we do not treat these applications here.

A granular material is a conglomeration of discrete solid particles (grains) ranging in size from 100 μm to the size of asteroids. Granular materials are ubiquitous in nature (sand, foam, snow, icebergs,...) and industry (drugs, cereals, sugar, fertilizer,...). They can behave like solids, liquids or gas [66] and they can present the paradoxical behaviour of segregation when one shakes a container with two different materials inside [55]. Mathematical models have been developed to study for example the dynamics of ice floes [153] or granular soils [49]. Many other models have been proposed for the different states of granular media mentioned above, we refer the reader to [15] for a review. The model and algorithms we propose in Chapter 3 could also be applied to the study of large granular systems.

In crowd dynamics, besides the presence of physical interactions, such as collisions, friction or clogging, each person is also driven by his/her own intentions and choices that influence the movement. Depending if the crowd is leaving a place in a normal situation or in an emergency, certain non-physical effects may occur, such as competition, collaboration, following behaviour, lane formation, crowd impatience, kin behaviour (for example, a family tries to stick together during evacuation), exit selection, etc. A big number of modelling approaches has been proposed in the literature to describe crowd dynamics during evacuations [143, 157] or during large sport, religious or cultural events [103, 110]. In particular, an agent based model has been proposed [124] where the contacts between agents are not anticipated and avoided, but they rather occur and are treated as non-smooth events. This is a realistic assumption in the case of congestion. The individual based model we present and explore throughout Chapters 1 to 4 follows the same principles and could therefore be used to study this type of systems.

0.3 Mathematical description of congested systems

We now consider mathematical models of the systems described in the previous Section. We start by presenting two classifications based on the spatial scale (particle based versus continuum models) and on the state of the system (statics versus dynamics), respectively (see Section 0.3.1). The remaining Sections are dedicated to an overview of mathematical and computational tools used in particle based models (Section 0.3.2) and continuum models (Section 0.3.3). In both Sections we start by presenting the class of optimization problems associated to the description of the statics equilibrium of the system (Sections 0.3.2.1 and 0.3.3.1), followed by a short literature review on models describing its evolution (Section 0.3.2.2 and 0.3.3.2).

0.3.1 Classification of models with volume exclusion constraints

0.3.1.1 Particle based versus continuum models

The models for congested systems can be divided into particle based* and continuum models.

Particle based models consider a geometric representation of individual particles and describe the evolution of the position and the state (for example, size, adhesion properties or tendency to behave in a certain way) of each particle under pre-defined forces acting on the particles. Particle based models provide information about the behaviour that occur at the particle scale. This may include the position, velocity, size and state of the particles over time, the rate of inter-particles interactions, the shape of particle aggregates, etc.

Continuum models describe averaged quantities such as number density and mean velocity, therefore they are applicable to systems with a large number of particles. Such quantities evolve in space and time under forces exerted on the particles. Continuum models constitute reduced descriptions of particle systems in which phenomena occurring at the particle level are averaged out. Instead, they provide information about macroscopic phenomena that is too difficult to be retrieved from a microscopic model, due to the large number of particles that that would require.

Particle based models for congested systems may be classified according to:

- space discretization: on-lattice or off-lattice models;
- the way congestion is modelled: encoded in the model, by considering a partition of space where each part represents one particle, such as vertex and Voronoi models, or as an emergent phenomenon, by considering particles that can move independently on their neighbours in between collisions;
- particle shape: simple or complex;
- the way contact events are modelled: non-smooth dynamics (hard-sphere) or approximated by a smooth dynamics (soft-sphere);
- stochastic or deterministic dynamics.

In Section 0.3.2.2 we use the criteria presented above to classify and compare several particle based models available in the literature as well as to motivate our own approach. Particularly, various factors need to be taken into account when making modelling choices. On the one hand, a complex model may be more realistic, but also more difficult to analyse.

*Particle based models (PBM) are also referred in the literature as individual based (IB) or agent based (AB) models. The difference may be related to the type of particle and the type of interactions between particles: PBM are typically used for systems of inert particles and IBM and ABM for systems of living particles. In this thesis we use PBM in general and IBM in the context of cell tissues.

On the other hand, a simple model may be easier to interpret, but may not be able to capture the main features of the system. Therefore, a right balance must be kept between complexity and simplicity.

Continuum models can be divided into two groups:

- phenomenological models: obtained from empirical observation of the system at the macroscopic scale as a fluid or a solid;
- coarse-grained models: derived from particle-based models via hydrodynamic or asymptotic limits.

This classification distinguishes two types of rules and modelling assumptions that one can use to build a continuum model: macroscopic and microscopic rules. In particular, the derivation of continuum models from particle-based models started with the well-known Hilbert's sixth problem, where he proposes developing "mathematically the limiting process (...) which leads from the atomistic view to the laws of motion of continua". Despite of the great advances during the last century, this problem continues to motivate researchers nowadays [44, 76]. In Section 0.3.3.2 we use this classification to present a short overview of continuum models and we situate our own approach.

This thesis is focused on the development and simulation of microscopic models for biological systems based on simple heuristic rules on the behaviour of the particles. This type of rules has also been used in the context of crowd motion [123]. In Chapter 5 we then derive a macroscopic model based on analogous heuristic rules to the ones of the microscopic model. This work intends to be a first step towards the connection between scales for the particular case of congested systems with contact events.

0.3.1.2 Statics versus dynamics

We divide the description of congested systems into statics (time-independent) and dynamics (time-dependent). In the first case, the system is assumed to be at *mechanical equilibrium*, corresponding to a state of minimal energy under volume exclusion constraints. The energy is associated to a given potential. In the second case, the energy and/or constraints evolve over time inducing *dynamics*. The system evolves by remaining at any time at mechanical equilibrium.

In this way, the state of the system at any time corresponds to a solution to a minimization problem with volume exclusion (or maximal density) constraints. In the setting we consider here, this leads to non-convex optimization in the particle based model and convex optimization in the continuum model as defined next.

Definition 1. *Let Y be a real vector space and f a real function on Y. If $\mathcal{C} \subseteq Y$ is a convex set and f is a convex function, then the minimization problem:*

find $x^* \in Y$ such that $x^* \in argmin\{f(x)|x \in \mathcal{C}\}$

is said to be **convex**, *otherwise, it is* **non-convex**.

In the particle based model, the non-overlapping constraints are described by inequalities which typically involve non-convex functions, leading to *non-convex optimization*. For example in the case of spherical particles, the non-overlapping constraint between two spheres of center x and y in $\mathbb{R}^d$, $d \in \mathbb{N}$ and radius R_1 and R_2, respectively, is given by $\phi(x, y) \leq 0$ with $\phi(x, y) = (R_1 + R_2)^2 - |x - y|^2$, where $|.|$ is the Euclidean norm (see Chapter 1). The function $\phi : (\mathbb{R}^d)^2 \mapsto \mathbb{R}$ is concave, and therefore the set of constraints is non-convex. When the energy function is a confinement potential, this problem can be seen as a sphere packing problem.

Contrarily, in the continuum model that we present in Chapter 5, the maximal packing constraints involve convex functions, leading to *convex optimization* (provided the confinement potential is also convex). Specifically, let $\tau : x \in \mathbb{R}^d \mapsto n(x) \in \mathbb{R}$ be a given function describing the average volume of particles in space and let $n : x \in \mathbb{R}^d \mapsto n(x) \in \mathbb{R}$ be the number density of particles over space, then the maximal density constraint can be written as $\phi(n) \leq 0$ with $\phi(n) = n\tau - 1$ (see Chapter 5).

In our continuum model, the convex problem can be solved analytically (see Chapter 5). Contrarily, in the microscopic case, there is no generic framework to tackle non-convex problems analytically, except in very particular cases. This motivates the use of iterative approaches to approximate a solution (see Chapters 1 and 2). In Sections 0.3.2.1 and 0.3.3.2 we provide the optimization tools in infinite and finite dimensional spaces needed throughout the thesis and we present in Section 0.3.2.1 a brief overview of the literature on algorithms for non-convex optimization.

The motion resulting from the evolution of the energy or the constraints may be smooth or non-smooth. For example, at the microscopic level, the contact events lead to discontinuities in the velocity of the particles, which makes the dynamics non-smooth. In Section 0.3.2.2 we discuss different approaches to model non-smooth events and we describe and classify the algorithms available for non-smooth dynamics. In contrast, at the macroscopic level the motion is typically smooth (see Chapter 5 and Section 0.3.3.2).

0.3.2 Particle based models and algorithms

0.3.2.1 Statics

In this Section we present the optimization tools needed to describe the static equilibrium of the system followed by a brief review on existing numerical methods for non-convex optimization. As explained in the previous Section 0.3.1.2, this state corresponds to the minimizer of an energy subject to non-overlapping constraints. Chapters 1 and 2 are dedicated to numerical solutions to this problem and they are based on a Lagrangian formulation which is presented here. In particular, we give the regularity conditions on the functions associated to the constraints that allow us to use this formulation.

Let p and q be positive integers. Consider the generic minimization problem with inequality constraints:

$$\begin{aligned} &\text{find } x \in \mathbb{R}^p \text{ such that} \\ &x \in \underset{x \in \mathcal{C}}{\operatorname{argmin}}\, W(x), \end{aligned} \tag{1}$$

with

$$\mathcal{C} := \{x \in \mathbb{R}^p \mid \phi_i(x) \leq 0,\ i = 1, ..., q\} \tag{2}$$

where $(\phi_i)_{i=1,...,q} = \phi : \mathbb{R}^p \to \mathbb{R}^q$ and $W : \mathbb{R}^p \to \mathbb{R}$ are given functions.

We define next the tangent and the linearized cone at a feasible point $\bar{x} \in \mathcal{C}$.

Definition 2. *The* ***tangent cone*** *of $\bar{x} \in \mathcal{C}$ is*

$$T(\bar{x}) := \{d \in \mathbb{R}^p \mid \exists \{x^n\}_n \subset \mathcal{C}, \{t^n\}_n \to 0^+ \ s.t. \ \ x^n \to \bar{x} \ \ and \ \frac{x^n - \bar{x}}{t^n} \to d\} \tag{3}$$

$$= \{d \in \mathbb{R}^p \mid \exists \{d^n\}_n \to d, \{t^n\}_n \to 0^+ \ s.t. \ \ \bar{x} + t^n d^n \in \mathcal{C},\ \forall n\}. \tag{4}$$

The ***linearized cone*** *of $\bar{x} \in \mathcal{C}$ is*

$$L(x) := \{d \in \mathbb{R}^p \mid \langle \nabla \phi_i(\bar{x}), d \rangle \leq 0,\ \forall i \ s.t.,\ \phi_i(\bar{x}) = 0\}, \tag{5}$$

where $\langle \cdot \rangle$ represents the inner-product in $\mathbb{R}^p$.

Note that if $\phi_i,\ i = 1, ..., q$ are continuously differentiable at $\bar{x}$, then one easily concludes that $T(\bar{x}) \subseteq L(\bar{x})$. We now introduce a regularity assumption on the functions associated to the constraints, also called constraint qualification.

Definition 3. *We say that the* **Abadie constraint qualification (ACQ)** *holds at $\bar{x}$ if $T(\bar{x}) = L(\bar{x})$.*

In Figure 4a we show an example where ACQ holds and in Figure 4b where it does not hold. Many other constraint qualifications have been proposed in the literature, most of them have been listed and compared in [148]. The next theorem ensures the existence of Lagrange multipliers under ACQ and provides the necessary conditions for optimality.

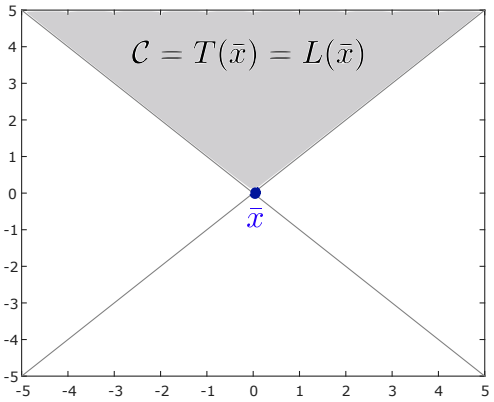
Theorem 4. *Let x^* be a solution to* (1) *such that the ACQ holds at x^*. Then there exist Lagrange multipliers, $\lambda \in \mathbb{R}^q$, such that the KKT-conditions hold, i.e.,*

$$\textit{feasibility conditions:} \quad \lambda_i \geq 0,\ \phi_i(x^*) \leq 0,\ i = 1, ..., q \tag{6}$$

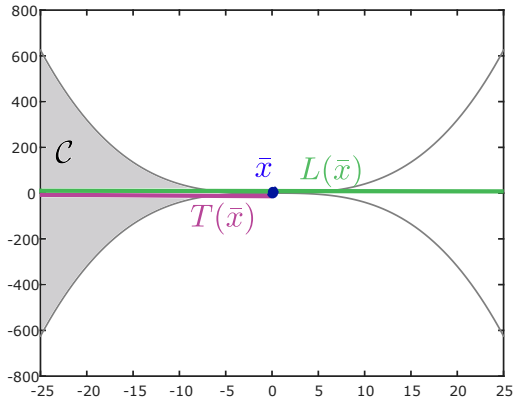
$$\textit{complementary slackness condition:} \quad \lambda \cdot \phi(x^*) = 0, \tag{7}$$

$$\textit{Euler-Lagrange equation:} \quad \nabla_x (W + \lambda \cdot \phi)(x^*) = 0, \tag{8}$$

where "$\cdot$" denotes inner-product.



(a) ACQ holds: $L(\bar{x}) = T(\bar{x})$



(b) ACQ does not hold: $L(\bar{x}) \neq T(\bar{x})$

Figure 4: Example (a) corresponds to $\mathcal{C} = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid -x_1 - x_2 \leq 0,\ x_1 - x_2 \leq 0\}$ and $\bar{x} = (0, 0)$. Example (b) corresponds to $\mathcal{C} = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid x_1^3 + x_2 \leq 0,\ x_1^3 - x_2 \leq 0\}$ and $\bar{x} = (0, 0)$.

Proof. See [2]. ∎

The conditions (6)-(8) are the so-called Karush-Kuhn-Tucker conditions (KKT - conditions) [134] and they are commonly used in the search for candidate solutions to the minimization problem (1). In Chapter 1 we develop an algorithm to approximate the solutions to this system of equations for the case of a non-convex minimization problem.

Classical algorithms to tackle non-convex problems (see definition 1) include Uzawa-Arrow-Hurwicz type algorithms [17], augmented Lagrangian [23, 25], linearly constrained Lagrangian (LCL), sequential quadratic programming (SQP) [134], among others. The SQP and the Uzawa-Arrow-Hurwicz algorithms are widely used. However they require the Hessian matrix of the function to be minimized to be positive definite, which is not always the case in this type of problems (see the example treated in Chapters 1 and 2). In general, all these methods perform well with a small number of particles. However we are interested in the case where this number becomes large. In Chapter 1 we propose a new algorithm that outperforms the classical ones when the number of particles is large. This algorithm is based on a modification of the Arrow-Hurwicz algorithm. We study its convergence numerically in Chapter 1 and analytically in Chapter 2 in the case of a specific problem involving a convex potential and non-convex constraints.

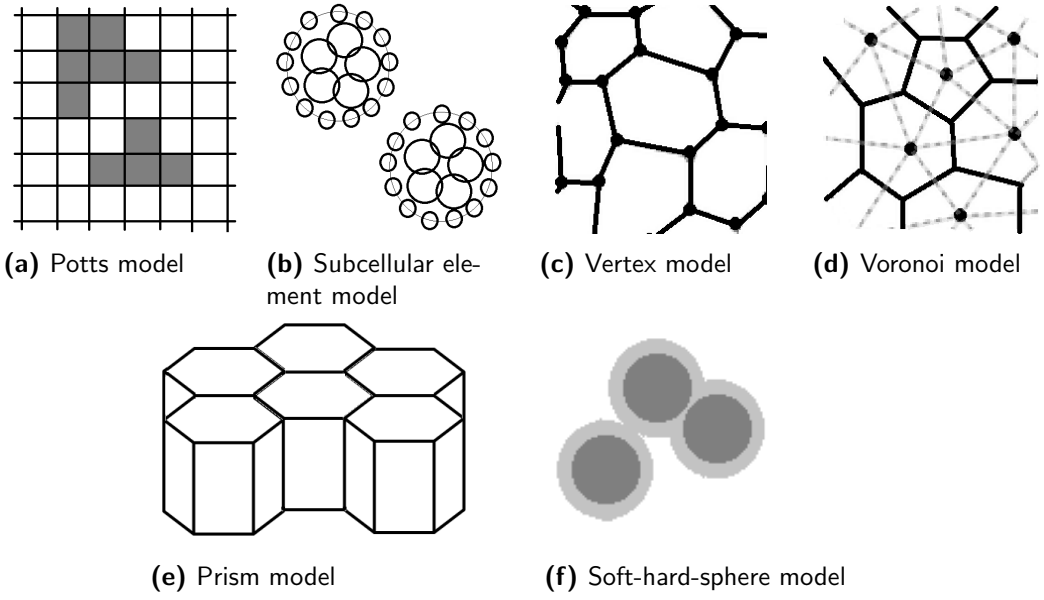
0.3.2.2 Dynamics

The simplest and perhaps the most popular discrete model to describe the dynamics of particle systems is the cellular automaton [179], where the space is discretized by a regular grid and a particle occupies one unit of the grid. A natural generalization of this model conduces to the Potts model, where a particle is constituted by a collection of units of the grid (see Figure 5a) and the evolution of the system is driven by the minimization of the total energy. Nevertheless, this model is able to account for complex shapes, which may

have contributed to its popularity in the modelling of the interactions between cells and with the environment, see for example [163]. These two methods use space discretization, which introduces grid artefacts, such as errors on the speed of the particles or on the inter-particle interactions [77, 126, 150]. It has been reported that the amount of artefacts increases with the particle density [126], which makes these models particularly unsuited to congested systems where the particle density is high.

Off-lattice models include for example the subcellular element model that has been used as a complex and more realistic description of cells [159]. A cell is represented as a collection of subcellular elements that interact with each other (see Figure 5b). This leads to a high degree of complexity that turns the study of large systems very challenging.

Other types of off-lattice models include vertex model and Voronoi model (see Figures 5c-5d, respectively). The former consists of a set of vertices connected through segments and has been used to model cell tissues, see for example [8, 59, 65, 168, 170]. The latter is based on a Voronoi tesselation (see [135] for a description of different types of tesselations and algorithms) and it has been used to model for example cell tissues [26, 127] and foams [167]. These models are suited to packed systems with no free space between the particles. The system is regarded as a partition of space where each part represents one cell.



(a) Potts model **(b)** Subcellular element model **(c)** Vertex model **(d)** Voronoi model

(e) Prism model **(f)** Soft-hard-sphere model

Figure 5: Discrete models for cell tissues. We use hard-spheres in the model for ballistic aggregation (see Chapter 3) and soft-hard-spheres in the model of a cell tissue (see Chapter 4).

In our modelling framework we would like to consider congestion as an emergent phenomenon, rather than directly encoded in the model, like in the vertex and Voronoi models. The most appropriate approach for our study is therefore a particle based model where each particle can move continuously in an off-lattice domain and collides with its neighbours. In order to describe the contact events, we need to describe the shape of the particle. To keep the model simple, we choose the simplest shape possible, which is a

hard-sphere. In Chapter 3 we use hard-spheres in the model for ballistic aggregation and in the Chapter 4 we use hard-spheres contained in soft-spheres to model cell nuclei in an epithelial tissue (see Figure 5f). Other approaches that also use hard-spheres have been used in [112, 149] in the context of cell tissues. Other simple particle shapes found in the literature are quadrilaterals [91] and hexagonal prisms [84] (see Figure 5e).

The contact events lead to discontinuities in the velocity of the colliding particles, which makes the dynamics intrinsically non-smooth. The remainder of this Section will be dedicated to the mathematical description and simulation of particle systems with contact events. They can be modelled as non-smooth events or be approximated by a smooth dynamics as explained next.

In the smooth approximation, the contact events do not occur instantaneously, they are rather solved during a short time-interval. This approach may be used to study large systems (with more than 10^4 particles) that do not require a high order of accuracy in the performance of contact events. One smooth approach to this problem is given by molecular dynamics algorithms, which typically consists of defining a (smooth) inter-particle potential and establishing and computing the Newton's equations for the motion of interacting particles [7]. Another smooth approach derives from the work by Cundall [41], which is very closely related to molecular dynamics except that it also accounts for rotations and complicated shapes.

In the non-smooth approach, the contact events are solved instantaneously at each time-step, which requires an extra computational effort. This approach is therefore more cumbersome, but also more accurate in the treatment of the volume exclusion constraints. In particular, it allows us to control the size of each particle at all times, which is crucial in the study of congestion. For this reason, in this thesis we opt for the non-smooth approach. More precisely, in the case of the model for a cell tissue developed in Chapter 4, we use both a hard and a soft sphere to allow some deformation and flexibility in the system and to make it more realistic. Specifically, we represent the nucleus of a cell by a hard-sphere that is contained in a soft-sphere, in this way, a nucleus may overlap other nuclei to a maximum point that is prescribed by the radius of the hard-sphere. This models the shrinking and squeezing of the nucleus at highly congested situations. Other non-smooth models have been presented in the literature, see for example [129] and [139, 140], where two different models of a mechanical system with unilateral contacts, impacts and Coulomb's friction are proposed in terms of the velocities and positions of the particles, respectively. Next we present an overview of numerical methods for simulating non-smooth dynamics.

The algorithms for non-smooth systems can essentially be divided into two classes which mainly differ in the way they deal with the non-smooth events: time-stepping (also called time-driven scheme) and event-driven [20, 111]. The former consists of defining *a priori* a regular and sufficiently fine discretization in time such that at each time-step all the events (such as collisions) that have occurred during the previous time-interval are

simultaneously solved (usually by integrating the Newton's second law and projecting onto the admissible space, which can be formulated as a minimization problem). The latter consists of computing the exact time of the next events, list them by chronological order and make the system evolving from one event to the next successively. The event-driven algorithms are usually more appropriate for systems with fewer events, since it is very likely that the events occur sequentially and not simultaneously. As the rate of events increases, the computation of the time of each event becomes more expensive and less accurate due to round-off errors [155]. At some point, an event-driven algorithm becomes too expensive computationally, therefore a time-stepping scheme needs to be used. In Chapter 3 we verify this effect in a model for ballistic aggregation with 10^6 particles, where the volume fraction of the particles is approximately $1/2$ of the domain. Since the occurrence of new collisions in a packed cell tissue is highly frequent, in Chapter 3 we opt for a time-stepping scheme in which our minimization algorithm is used at each time-step to solve the events that occurred during the previous time-interval. For a very complete overview on numerical methods for non-smooth dynamics problems see [28]. We next provide a brief review on time-stepping algorithms.

There are two main time-stepping schemes available to integrate nonsmooth dynamics [3]: the Moreau-Jean scheme [129, 161] and the Schatzman-Paoli scheme [139, 140]. These schemes have been proposed for systems with unilateral contacts, impacts and Coulomb's friction. The main difference between the two schemes is that in the former the constraints are imposed at the velocity level, while in the latter they are imposed at the positions level. The idea is the following: at each time-step, the new positions and velocities of the particles are computed according to the dynamics of the system, and an eventual violation of the constraints (non-admissible configuration) is then corrected by projecting the velocities (resp. positions) onto the set of admissible velocities (resp. positions). This corresponds to solving a minimization problem where one minimizes the distance between the non-admissible velocities (resp. positions) and the set of admissible velocities (resp. positions) as follows: given the velocities $\mathbf{V}_0$ (resp. positions $\mathbf{X}_0$) computed by the dynamics of the system, find the admissible velocities $\bar{\mathbf{V}}$ (resp. positions $\bar{\mathbf{X}}$) that satisfy

$$\bar{\mathbf{V}} \in \operatorname{argmin}_{\mathbf{V}\in\mathcal{C}} \operatorname{dist}(\mathbf{V}_0, \mathbf{V}) \qquad (\text{resp. } \bar{\mathbf{X}} \in \operatorname{argmin}_{\mathbf{X}\in\mathcal{C}} \operatorname{dist}(\mathbf{X}_0, \mathbf{X})), \tag{9}$$

where $\mathcal{C}$ is the admissible set of velocities (resp. positions) and dist is a distance. This operation may introduce long range nonlocal effects. In Chapter 3 we consider a simple experimental set up which shall be used in the future to investigate how these effects may propagate in time and space by comparison with an exact event-driven algorithm.

We deduce from the description presented in the previous paragraph that in Moreau-Jean scheme, the collision rule is strictly satisfied, but not the constraints in position, and it is the opposite for the Schatzman-Paoli scheme. Moreover the multiplier associated to the constraints may be directly interpreted as a mechanical impulse in the Moreau-Jean

scheme, while in the Schatzman-Paoli it does not have any direct mechanical interpretation. In [3] the author proposes a directed projection scheme that satisfies in discrete time both the position and velocity constraints.

In our numerical experiments in Chapter 1 and 3, we observe that when the constraints are imposed onto the velocities, the computational cost of the minimization algorithm becomes highly dependent on the initial configuration. This becomes a especially serious problem when the number of particles is high (see Chapter 1). We therefore follow an approach that is similar to the one by Schatzman-Paoli scheme, except that instead of computing the projection at each time-step, we compute a minimizer of a given energy function subject to non-overlapping constraints using our minimization algorithm. The case where the energy function is the distance between the variable $\mathbf{X}$ and the current iterate $\mathbf{X}_0$, as in the minimization problem (9), corresponds to the Schatzman-Paoli scheme. In this sense, our approach is more general.

The other type of algorithms for non-smooth systems, the event-driven algorithms, are less common. This may be due to the high computational cost they require, but also to the difficulty on analysing their convergence, which may become a problem in the simulation of complex dynamics [28]. Event-driven algorithms have been used for instance in [152, 153] to study the dynamics of ice floes, in [5] to study rheology of granular mixtures and in [51] to study jammed packings of spheres and ellipsoids.

In Chapter 3, we consider a simple 2D system constituted of hard-circles undergoing ballistic aggregation in a periodic domain. We use an event-driven algorithm as follows: we compute the exact time of the next collision and we perform the collision exactly. This algorithm allows us to get the exact dynamics of the system (apart from round-off errors). Despite being more accurate, the event-driven algorithm takes much longer to converge when the number of particles and the rate of events increases (see Chapter 3), which leads us to favour time-stepping schemes for the simulation of congested systems.

In Figure 6 we present a classification of the main time-stepping and event-driven algorithms that have been proposed in the literature for smooth and non-smooth dynamics. The algorithms we develop in this thesis belong to the class of time-stepping schemes and event-driven for non-smooth dynamics (see Chapter 3).

0.3.3 Continuum models

0.3.3.1 Statics

In this Section we present the tools needed to describe the static equilibrium of the system. As explained in Section 0.3.1.2, this state corresponds to the minimizer of an energy subject to maximal density (packing) constraints. Contrarily to the minimization problem arising in the discrete model (see Section 0.3.2.1), in the continuous model the minimization problem is convex and we are able to solve it analytically. The resolution

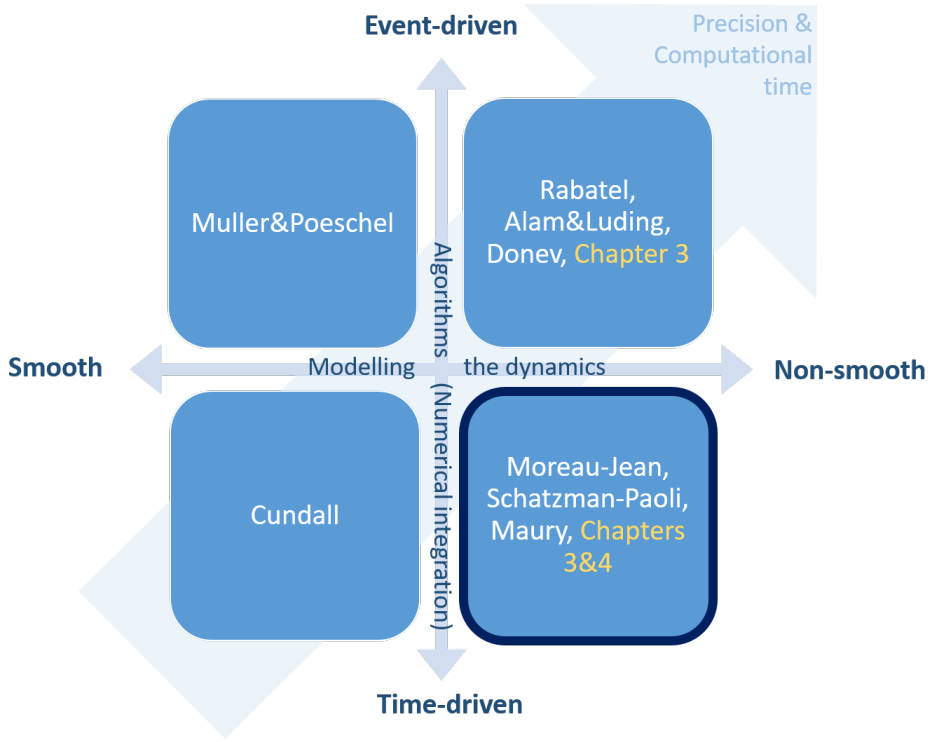


Figure 6: Classification of the algorithms for smooth and non-smooth dynamics of many particle systems with volume exclusion constraints. Our approach belongs to the class of time-stepping schemes for non-smooth dynamics.

is presented in Chapter 5 and it uses the Lagrangian formulation for minimization problems in infinitely dimensional spaces that we present next. In particular, we provide the regularity conditions on the functions associated to the constraints that allow us to use this formulation.

Let Y be a real vector space. Consider the minimization problem

$$\begin{aligned} &\text{find } n \in Y \text{a solution of} \\ \min\{f(n) \quad &| n \in Y,\ g_i(n) \leq 0,\ i = 1, ..., p,\ h(n) = 0\}, \end{aligned} \tag{10}$$

where $f : Y \to \mathbb{R}$, $(g_i)_{i=1,...,p} = g : Y \to \mathbb{R}^p$ and $(h_j)_{j=1,...,q} = h : Y \to \mathbb{R}^q$ are the given data. If f and $g_i,\ i = 1, ..., p$ are convex and $h_j,\ j = 1, ..., q$ are affine functions then by definition 1 the minimization problem is convex.

Consider the Slater condition:

- there exists a strictly admissible point n_0 for (10), i.e.,

$$g(n_0) < 0 \quad \text{and } h(n_0) = 0,$$

- the equality constraints are independent,

and the regularity assumptions:

- f, g and h are continuously (Fréchet) differentiable in a neighbourhood of n^*,
- the minimization problem is convex and f, g and h are continuous and Gateaux differentiable on Y.

The next theorem ensures that the necessary conditions for optimality hold true provided n^* is a solution to the minimization problem and some regularity assumptions are satisfied. As we shall see, if the problem is convex the continuous differentiability assumption can be relaxed.

Theorem 5. *Let n^* be a solution to (10). Suppose the Slater condition holds and at least one regularity assumption holds. Then there exist multipliers $(\lambda,\ \mu) \in \mathbb{R}^p \times \mathbb{R}^q$ satisfying*

$$\textit{non-triviality conditions:} \quad (\lambda, \mu) \neq 0, \tag{11}$$

$$\textit{feasibility conditions:} \quad \lambda_i \geq 0,\ g_i(n^*) \leq 0,\ h(n) = 0,\ i = 1, ..., p, \tag{12}$$

$$\textit{complementary slackness condition:} \quad \lambda \cdot g(n^*) = 0 \tag{13}$$

$$\textit{Euler-Lagrange equation:} \quad (f + \lambda \cdot g + \mu \cdot h)'(n^*) = 0, \tag{14}$$

where "′" denotes Fréchet derivative and "· " denotes inner-product.

Proof. See Thm. 9.4 and Thm. 9.8, p. 182 – 185 in [35]. ∎

This is a central result in optimization and is used in this thesis in Chapter 5.

0.3.3.2 Dynamics

We now present a brief overview on continuum models for congested systems using the classification proposed in Section 0.3.1.1, which divides the models into two groups: phenomenological models and coarse-grained models (which are derived from particle-based models). Within each group they can be subdivided according to their underlying framework into reaction-diffusion systems, solid mechanics, fluid mechanics and free-boundary problems.

We first present a few examples of phenomenological models and the contexts where they have been applied.

- *Reaction-diffusion systems.* Reaction-diffusion systems have been used in the study of tumour spheroids [162] and invasion processes applied to cell biology and ecology [64, 104, 132].
- *Solid mechanics.* We refer to [11] where a model of swelling gel is proposed in the framework of hyperelasticity theory and applied to the study of certain tumours, such as malignant melanoma. In the context of tumor growth modelling a solid mechanics model can be found for example in [32].

- *Fluid mechanics.* A fluid model has been proposed in the context of crowd dynamics to study the transition between congested and uncongested regimes [46]. In the context of tumour growth, fluid models based on Darcy's law can be found for example in [22, 31, 37, 40, 87].

- *Free-boundary problems.* Free-boundary problems are used in the modelling of tissue and tumour growth. They describe the boundary of the tissue over time [71]. In [21, 90] the authors have explored under which asymptotic limits the fluid model could lead to a free-boundary problem. Related to these, the analogy between tumor growth and the free-boundary problem of solidification (the so-called Hele-Shaw problem) has been developed in [88, 145, 146, 147].

We now present a few examples of continuum models that have been derived from particle based models (see Section 0.3.2.2 for a description of the different types of particle based models).

- *Reaction-diffusion system derived from a lattice-based model.* Reaction-diffusion systems for cell migration and proliferation have been derived from lattice-based models through a mean-field limit in one [96] and two [95] spatial dimensions.

- *Model in solid mechanics derived from vertex model.* Also in the context of cell biology, an epithelial cell tissue has been described by a two-dimensional viscoelastic sheet originated from a vertex model through coarse-graining [8].

- *Fluid model derived from off-lattice model with soft-spheres.* A fluid model describing the dynamics of herds [47] has been derived from an off-lattice model with soft non-overlapping constraints through mean-field and hydrodynamic limits.

- *Free-boundary problem derived from off-lattice model with hard-spheres.* In the context of tumour growth, a Hele-Shaw type problem has been formally derived through coarse-graining from an off-lattice particle model with hard-contraints [130].

In Chapter 5 we consider a continuum model based on the minimization of a confining potential under maximum packing density constraint. We then prove that these rules give rise to a free-boundary problem. Similarly to the Hele-Shaw problem, the system is regarded as the region of space where particles have reached the packing density. The system is therefore an incompressible medium separated from the outer medium by a moving free boundary. Since the heuristic rules considered in the continuum model are analogous to the ones considered in the particle-based model, this work intends to be a first step towards a formal (and rigorous) derivation of the continuum model from the discrete one.

0.4 Aims, main results and contributions

The thesis is divided into three parts. Part I deals with algorithms for static packing problems, which represent a fundamental tool in the discrete modelling framework presented in Part II. Part II is mostly concerned with algorithms simulating dynamical problems and in particular the study of an epithelial tissue. In Part III we develop a continuum framework for swelling materials. The main results are summarized next.

0.4.1 Algorithms for packing problems (Part I)

We present next the general problem that motivates the work discussed in Part I of this thesis. Let $d \in \mathbb{N}$ be the spatial dimension and $N \in \mathbb{N}$ the number of particles in $\mathbb{R}^d$, with position $\mathbf{X} = \{X_i\}_{i=1,\dots,N}$. Consider the following *local* minimization problem: find a configuration $\bar{\mathbf{X}} \in \mathbb{R}^{Nd}$ such that

$$\bar{\mathbf{X}} \in \underset{\phi_{k\ell}(\mathbf{X}) \leq 0,\ k,\ell=1,\dots,N,\ k<\ell,\ \mathbf{X} \in U(\bar{\mathbf{X}})}{\operatorname{argmin}} W(\mathbf{X}), \tag{15}$$

where $U(\bar{\mathbf{X}})$ is a neighbourhood of $\bar{\mathbf{X}}$, $W : \mathbb{R}^{dN} \to \mathbb{R}$ is a convex and smooth function (not necessarily strictly convex) with an attainable minimum in the admissible set defined by the constraints. The functions $\phi_{k\ell} : \mathbb{R}^{dN} \to \mathbb{R},\ k < \ell$ are continuous but not necessarily convex. In these conditions, $\bar{X}$ exists, but may not be unique. Note that we are not necessarily interested in finding a global minimizer. Indeed, in many systems the particles only seek to achieve a locally optimal solution. To simplify the notation we will omit the condition $\mathbf{X} \in U(\bar{\mathbf{X}})$ in the formulation of the minimization problems presented throughout the thesis.

Specifically, we focus on the case where W represents a *confining potential* and $\phi_{k\ell}$ are associated to *non-overlapping constraints* between particles k and ℓ. In this setting, the minimization problem (15) can be seen as a *packing problem*. Note that the functions associated to non-overlapping constraints are typically *non-convex*, which brings additional difficulties compared to convex problems (see Section 0.4.1.1).

Examples of possible solutions in dimensions $d = 2$ and $d = 3$ are shown in Figure 7, with

$$\phi_{k\ell}(\mathbf{X}) = (2R)^2 - |X_k - X_\ell|^2, \tag{16}$$

and

$$W(\mathbf{X}) = \frac{1}{2N} \sum_{i<j} |X_i - X_j|^2. \tag{17}$$

where R denotes the radius of the spheres and $|.|$ the Euclidean norm.

In Chapter 1 we present a new iterative algorithm to find an approximate solution to (15) and we explore its computational speed and accuracy compared with classical

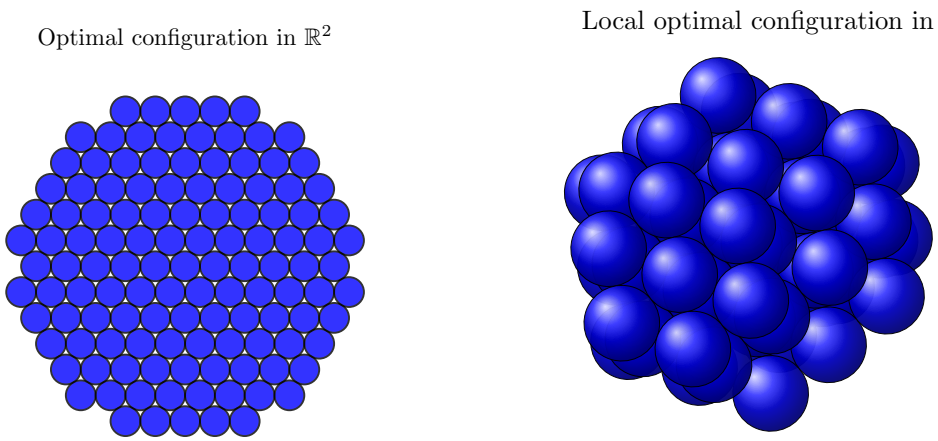


Figure 7: Two examples of solutions to the packing problem (15) with $\phi_{k\ell}$ and W defined in (16) and (17), respectively. A global optimum in dimension 2 is presented on the left and corresponds to the hexagonal lattice. A local optimum in dimension 3 is presented on the right.

methods in the case of a sphere packing problem with W defined by (17). In Chapter 2 we consider again the confining potential (17). We prove the convergence of a first-order formulation of the algorithm towards the theoretical solution in dimension $d = 1$ and we develop the analysis of speed of convergence in any dimension.

0.4.1.1 Damped Arrow-Hurwicz algorithm for sphere packing (Chapter 1)

Chapter 1 is based on the published work [45] which is a collaboration with Pierre Degond and Sebastien Motsch. The aims of this Chapter are:

- To develop an efficient and accurate algorithm to solve the non-convex minimization problem (15).
- To test numerically the new algorithm against other algorithms existing in the literature.

We start by reducing the original problem (15) to a critical-point system with a possible enlargement of the set of solutions. Under certain regularity conditions on the functions $\phi_{k\ell}$, from theorem 4 (Section 0.3.2.1) we have that, if $\bar{\mathbf{X}}$ is a solution to the minimization problem (15), then, there exists $\bar{\boldsymbol{\lambda}} \in (\mathbb{R}_0^+)^{N(N-1)/2}$ such that $(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}})$ is a critical-point of the Lagrangian $\mathcal{L} : \mathbb{R}^{dN} \times (\mathbb{R}_0^+)^{N(N-1)/2} \to \mathbb{R}$ defined by

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}) = W(\mathbf{X}) + \sum_{k,\ell \in \{1,\dots,N\},\ k<\ell} \lambda_{k\ell} \phi_{k\ell}(\mathbf{X}),$$

where $\boldsymbol{\lambda} = \{\lambda_{k\ell}\}_{k,\ell=1,\dots,N,\ k<\ell}$ represents the set of Lagrange multipliers.

In the case where the functions $\phi_{k\ell}$ are non-convex, there may be critical-points that are not solutions to the original minimization problem. Moreover, contrarily to convex optimization, a critical-point may not be a saddle-point (see Section 2.4.1). This is the

starting point of our work. We first observe that the classical Uzawa-**A**rrow-**H**urwicz **A**lgorithm (AHA) does not converge. Specifically, by writing the algorithm as the discretization of an ODE system, we realize that the non-existence of saddle-points leads to the existence of periodic orbits around the steady-state. To overcome this problem, we add a damping term to the system of ODEs in order to change its dynamics without changing the steady states. By discretizing the modified ODE system we obtain what we call the damped Arrow-Hurwicz algorithm (DAHA) that is presented in the next definition.

Definition 6. *We define the* ***damped Arrow-Hurwicz system (DAHS)*** *as*

$$\begin{cases} \ddot{X}_i &= -\alpha^2[\nabla_{X_i} W(\mathbf{X}) + \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X})] \\ &\quad -\alpha\beta \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X})\lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X}) - c\dot{X}_i\ , \\ &\qquad\qquad\qquad\qquad\qquad\qquad i = 1,\dots,N \\ \dot{\lambda}_{k\ell} &= \begin{cases} 0, & \text{if } \lambda_{k\ell} = 0 \text{ and } \phi_{k\ell}(\mathbf{X}) < 0 \\ \beta\phi_{k\ell}(\mathbf{X}), & \text{otherwise} \end{cases}, \\ &\qquad\qquad\qquad\qquad\qquad\qquad k,\ell = 1,\dots,N,\ k<\ell \end{cases}$$

where α, β and c are positive constants and the ***damped Arrow-Hurwicz algorithm (DAHA)*** *as the corresponding semi-implicit discrete scheme:*

$$\begin{cases} X_i^{n+1} &= \frac{1}{1+c/2}\left(2X_i^n - (1-c/2)X_i^{n-1}\right) \\ &\quad -\frac{\alpha^2}{1+c/2}[\nabla_{X_i} W(\mathbf{X}^n) + \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}^n\nabla_{X_i}\phi_{k\ell}(\mathbf{X}^n)] \\ &\quad -\frac{\alpha\beta}{1+c/2} \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X}^n)\lambda_{k\ell}^n\nabla_{X_i}\phi_{k\ell}(\mathbf{X}^n), \\ &\qquad\qquad\qquad\qquad\qquad\qquad i = 1,\dots,N \\ \lambda_{k\ell}^{n+1} &= \max\{0, \lambda_{k\ell}^n + \beta\phi_{k\ell}(\mathbf{X}^{n+1})\},\ k,\ell = 1,\dots,N,\ k<\ell, \end{cases}$$

where α, β and c correspond now to numerical parameters.

We prove that the algorithm converges to a solution of our original problem for the particular potential (17) with $N = 2$ in one spatial dimension and we observe numerically that it also converges in two spatial dimensions for a large number of spheres.

Our algorithm is then compared with two different methods belonging to the class of linearly constrained Lagrangian methods (LCL). The first method consists of solving a sequence of convex minimization problems with linearized constraints using the AHA. At each iteration, the constraints are linearized around the solution obtained at the previous iteration. The numerical implementation involves an outer and an inner loop, what will be referred to as the nested algorithm for the positions (NAP). The second method is based on a formulation of the problem involving the evolution of $\nabla_{X_i} W(\mathbf{X}),\ i = 1,\dots,N$ which we will refer to as the velocities. More precisely, the problem is formulated iteratively: given an initial configuration $\mathbf{X}^0 = \{X_i^0\}_{i=1,\dots,N}$, we pass from iterate $\mathbf{X}^p$ to iterate $\mathbf{X}^{p+1}$

as follows

$$\begin{cases} \mathbf{X}^{p+1} = \mathbf{X}^p + \tau\ \mathbf{V}^{p+1} & (18) \\ \mathbf{V}^{p+1} \in \underset{\phi_{k\ell}(\mathbf{X}^p + \tau\ \mathbf{V}) \leq 0,\ k,\ell=1,\dots,N,\ k<\ell}{\operatorname{argmin}} \dfrac{1}{2} \displaystyle\sum_{i=1}^{N} |V_i + \nabla_{X_i} W(\mathbf{X}^p)|^2, & (19) \end{cases}$$

where $\tau > 0$ is a given parameter and $\mathbf{V} = \{V_i\}_{i=1,\dots,N}$. Similarly to the first method, we use the AHA to solve the minimization problem (19) with linearized constraints. This algorithm will be referred to as nested algorithm for the velocities (NAV).

Numerical experiments in dimension $d = 2$ were designed to assess the three algorithms (DAHA, NAP, NAV) in a scenario where particles are strongly attracted to each other as exemplified by the global potential (17) and a highly dense and overlapping initial configuration. After spanning the parameter space, the best parameters for each method were selected. In a second stage, the three methods were compared in terms of the computational time, accuracy and variability of computational time with respect to initial configurations. The NAP turned out to be the fastest, the most accurate and the less dependent on initial conditions for a small number of spheres ($N = 7$). For a large number of spheres ($N = 100$), the DAHA performed better than any other method in terms of computational time and variability of computational time. Accuracy has not been measured as for N large the solution of the minimization problem is not unique.

Finally, in dimension $d = 3$, we show that the DAHA is able to generate random close packings with a volume fraction of approximately 0.65. Whether a higher fraction could be achieved by choosing a different set of parameters remains an open problem.

Future directions include the analysis of numerical stability and the study of the performance of the algorithms for other types of initial configurations, higher spatial dim and other particle shapes. Some of the numerical results obtained in this work are analytically investigated in Chapter 2.

The applications of particle systems with volume exclusion constraints are many and we believe that these type of algorithms may be very useful for biological, physical and social systems.

0.4.1.2 Analytical study of a non-convex minimization problem (Chapter 2)

Chapter 2 is based on a collaboration with Sebastien Motsch. The aims of this Chapter are:

- To develop an analytical study of the new algorithm proposed in Chapter 1 in terms of convergence and speed of convergence in 1D and for an arbitrary number of particles.
- To compare the new algorithm with an augmented Lagrangian method in terms of

speed of convergence.

We consider the non-convex minimization problem (15) with the global potential (17). We start by studying properties of a solution to this problem itself. In particular, we fully characterize the only solution (apart from permutations and translations) and the associated Lagrange multipliers in dimension $d = 1$ for any value of N. In dimension $d > 1$, there are in general many local minima and there is no explicit expression for the Lagrange multipliers. Nevertheless, we are able to compute the order of magnitude with respect to N of the Lagrange multipliers associated to a particular class of solutions belonging to the set of spherical configurations.

We then analyse the dynamical system associated to the AHA and prove that it does not converge. Specifically, we show that for $N > d+1$ the Lagrangian does not have any saddle-point.

We dedicate the remaining of Chapter 2 to the analysis of two dynamical systems. The first one is associated with an augmented Lagrangian algorithm (ALA) and we call it augmented Lagrangian system (ALS). The second one is associated with a first-order formulation of the DAHS and we call it 1st-order DAHS. The two systems have the form:

$$\begin{cases} \dot{X}_i = -\alpha[\nabla_{X_i} W(\mathbf{X}) + \sum\limits_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X})] \\ \qquad\quad -\gamma D_i(\mathbf{X},\lambda), \\ \qquad\qquad\qquad\qquad i = 1,\dots,N \qquad (20) \\ \dot{\lambda}_{k\ell} = \begin{cases} 0, & \text{if } \lambda_{k\ell} = 0 \text{ and } \phi_{k\ell}(\mathbf{X}) < 0 \\ \beta\phi_{k\ell}(\mathbf{X}), & \text{otherwise} \end{cases}, \\ \qquad\qquad\qquad\qquad k,\ell = 1,\dots,N,\ k<\ell, \qquad (21) \end{cases}$$

where γ and β are positive constants. In the ALS, the last term in equation (20) is defined by

$$D_i(\mathbf{X},\lambda) := D_i^{AL}(\mathbf{X},\lambda) = \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X})\nabla_{X_i}\phi_{k\ell}(\mathbf{X})$$

and in the 1st-order DAHS the same term is defined by

$$D_i(\mathbf{X},\lambda) := D_i^{DAH}(\mathbf{X},\lambda) = \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X})\lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X}).$$

We prove the convergence of both systems in dimension $d = 1$ towards the solution of the minimization problem characterized before, provided the initial configuration is sufficiently close to the steady state. Specifically, we use linear stability analysis to prove that the solution to the minimization problem is an asymptotically stable steady state of (20)-(21) for any value of the parameters. Moreover, we obtain for each system the region in the parameter space where the convergence is faster. We conclude that in the case of the 1st-order DAHS this region is narrower but it may lead to a faster convergence

than the corresponding region of the ALS.

There is still much more work left to do, namely, investigate the convergence properties of the DAHA towards a solution to the minimization problem in high spacial dimensions, finding a rule to choose the parameters that lead to a faster convergence and study the numerical stability of the algorithms. Despite of the simplicity of the formulation and resolution of sphere packing problems in low dimensions, they turn out to be extremely complex in higher dimensions (above 2). Therefore they still remain nowadays an active area of research in mathematics [36, 89].

0.4.2 Particle-based models for congested systems (Part II)

We now consider discrete descriptions of the time-evolution of particle systems with congestion. At each time, we assume that the system is at a mechanical equilibrium. Such equilibrium state corresponds to a solution of a packing problem (see Part I), namely, a minimization of a potential energy (confinement potential) subject to volume exclusion constraints. The evolution of the system is then driven by changes in the potential energy and other factors that are specified in the next Sections. This part of the thesis is formed of two Chapters:

- In Chapter 3 we present a numerical study on the performance of several algorithms for simulating the dynamics of congested systems. The experimental set up is a particle-based model for ballistic aggregation. We compare several event-driven and time-stepping algorithms and we select the one that performs faster for a large number of particles.

- In Chapter 4, we build a mechanical model of a packed cell tissue and we use the algorithm selected in Chapter 3 to simulate its dynamics. Using the model, we then show that some hypotheses about the dynamics of the tissue are consistent with the real data.

0.4.2.1 Event-driven versus time-stepping schemes for ballistic aggregation (Chapter 3)

Chapter 3 is based on a collaboration with Pierre Degond, Giacomo Dimarco and Sophie Hecht. The aims of this Chapter are:

- To develop a time-stepping algorithm that efficiently simulates the dynamics of large and dense particle systems.

- To assess the new algorithm against an event-driven algorithm in terms of computational time in the context of a particle system undergoing ballistic aggregation.

Consider $N \in \mathbb{N}$ self-propelled hard-circles, with radius $R > 0$ and positions $\mathbf{X} = \{X_i\}_{i=1,\dots,N} \in \mathbb{R}^{2N}$. The circles move smoothly in straight trajectories in a torus until they collide and stick, forming growing clusters with velocity equal to the average of the velocities of the constituent particles. We consider the evolution of the system until all particles have come together to form one big cluster. The phenomenon described is called *ballistic aggregation* and constitutes our experimental set up to assess an event-driven (ED) and four time-stepping (TS) schemes (see Section 0.3.2.2 for a description of the main features of these methods). The goal is to find a scheme that is able to deal with a large number and a large density of particles.

In the TS schemes we define a small time-interval Δt. Given the state of the system at time t, we compute the state at time $t+\Delta t$ by performing three steps: 1) the particles move at a constant velocity until time $t + \Delta t$, 2) the positions of the particles are adjusted to restore the non-overlapping condition in the system, 3) the velocities of colliding clusters are updated to the average velocity of the constituent particles. Step 2) consists of solving a minimization problem with non-overlapping constraints (see Part I). The four different TS schemes (S-R, NS-R, S-NR and NS-NR) are obtained by combining the two following aspects:

- rigid (R) or non-rigid (NR) clusters, namely, each cluster behaves either as a whole or as a collection of particles. In the former, all particles of the same cluster are subject to the same update during the minimization algorithm and in the latter each particle is updated individually;

- the use of smooth (S) or non-smooth (NS) functions associated to the non-overlapping constraints, $\phi_{k\ell}(\mathbf{X}) \leq 0,\ k < \ell$, namely,

$$\phi_{k\ell}(\mathbf{X}) = (2R)^2 - |X_k - X_\ell|^2$$

  or

$$\phi_{k\ell}(\mathbf{X}) = 2R - |X_k - X_\ell|,$$

  respectively. Note that the two functions lead to equivalent constraints, nevertheless they have different influence on the speed and accuracy of the minimization algorithm.

We then compare the computational time of the TS and ED for different numbers of spheres N and different sizes of the domain. We observe that when the number of particles becomes large or the size of the domain becomes small, the ED method becomes computationally more expensive than the TS schemes. Precisely, in our experiments, we observe that for for $N = 10^6$ circles and a domain $[0, L]^2$, with $L = 2.5R\sqrt{N}$, the ED takes almost 4 days, while the TS takes less than two hours[†].When comparing the

[†]The computational time was obtained in a Dell PowerEdge R630 rackmount server fitted with two Intel Xeon E5-2637 3.5 GHz processors, each of which with 4 cores and 8 threads.

TS algorithms between themselves, we observe that the ones using the smooth form of the constraints are faster than the corresponding methods with non-smooth form of the constraints. In Chapter 4 we therefore opted by the smooth form of the constraints in the model for a cell tissue.

In future work, the TS methods should be assessed in terms of how accurate they represent the real dynamics. Since the ED scheme is an exact method, we could use it in this assessment as the reference for accuracy. After selecting the most efficient and accurate TS scheme, we could use it to study statistical properties associated to ballistic aggregation. By introducing variability in the initial positions and velocities, one could explore the evolution of the number of clusters over time or the type of shapes arising during and at the end of the dynamics.

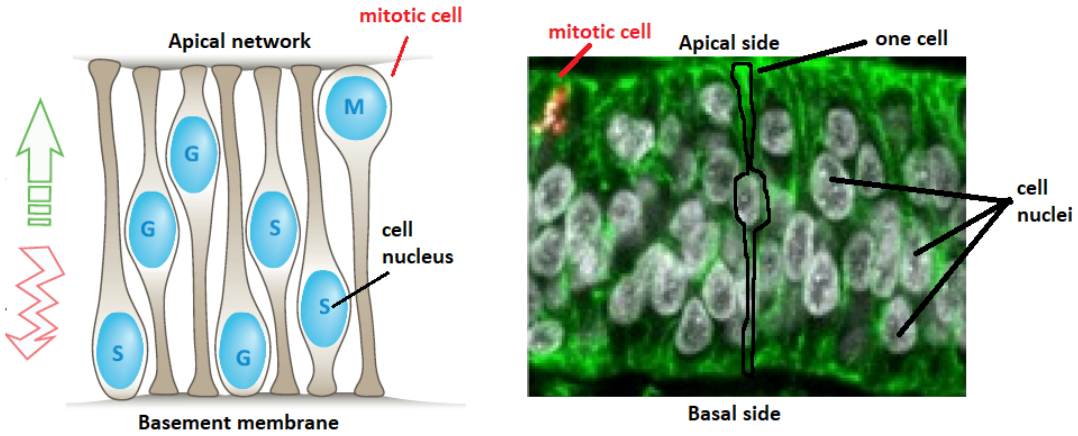
0.4.2.2 Modelling a packed cell tissue with noise (Chapter 4)

Chapter 4 is based on a collaboration between mathematicians Pierre Degond and Sara Merino-Aceituno and a team of biologists from Toulouse led by Eric Theveneau and including Fernando Duarte. The aims of this Chapter are:

- To develop a mechanical model for a packed cell tissue.
- To explore the parameter space of the model.
- To obtain a qualitative and quantitative comparison with real data.
- To come up with preliminary hypotheses about the mechanics of the tissue.

In a densely packed tissue the contact forces between neighbouring cells cannot be neglected. We consider in particular tissues where the nucleus is the most rigid part of the cell, leading to nuclei packing. Our goal is to understand the contribution of the mechanical interactions between neighbouring cells to the shape and dynamics of the tissue. In order to address this question, we compare experimental results with simulations of a mathematical model. The experiments conducted by the team of Eric Theveneau are performed on the chick neuroepithelium that will be presented next. The aim of the mathematical model is to test hypotheses that are based on the observation of experimental data.

The neuroepithelium is constituted by one layer of elongated column-shaped cells whose nuclei can be found at different heights inside the cell, giving the perception of a multilayer tissue (see Figure 8b). The cells are involved by an extra-cellular matrix. Epithelial cells have a polarity, at the bottom (basal side), the cells are attached to a membrane, called basement membrane, and at the top (apical side), they are attached to each other, forming the so-called apical network (see Figure 8a). The nucleus moves inside the cell along the apico-basal axis. Prior to division, it migrates to the apical side, where it



(a) Sketch of the organization of a pseudo-stratified epithelial tissue. Image taken from [109] and modified.

(b) Image of the chick neuroepithelium taken by Theveneau's team.

Figure 8: In a pseudo-stratified epithelial tissue the nuclei are found at different heights inside the cells. Prior to cell division, the nucleus migrates inside the cell to the apical side of the tissue.

divides. How the nucleus migrates and how that relates to the shape of the tissue are questions we want to address in this study.

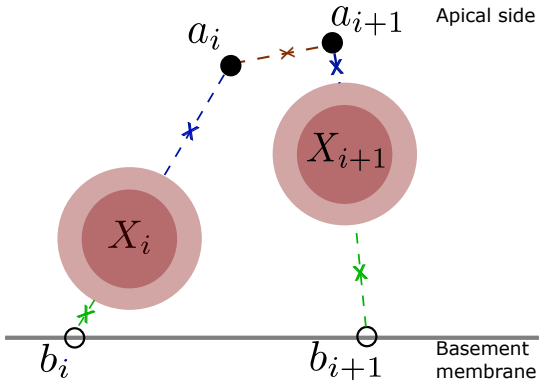


Figure 9: Sketch of the model for the tissue presented in Figure 8. The model consists of soft-hard-spheres connected to the apical and basal points through springs. The basal points are restricted to a straight line and they cannot switch positions nor get too far away from each other. Neighbouring apical points are connected through springs. The apical point, nucleus and basal point are subject to an alignment force.

In the mathematical model, a cell is constituted by the nucleus, an apical point and a basal point. The apical and basal points are connected to the nucleus through springs with adaptive rest lengths (see Figure 9). The rest length is gradually adapting to the actual length of the spring, which introduces a visco-elastic effect on the system. The nucleus is modelled by an inner-circle, which represents an impenetrable hard-core and an outer-circle that resists to compression through an elastic response. The basal points are restricted to the basement membrane, which is represented by a straight line. Moreover, they cannot switch positions nor get too far away from each other, which models lateral adhesion, i.e., the fact that cells adhere to each other. Each apical point is connected to its neighbouring apical points through springs, forming an apical network. We also consider an alignment force that tends to align the apical point, the nucleus and the basal

point, which models the influence of the cytoplasm of the cell.

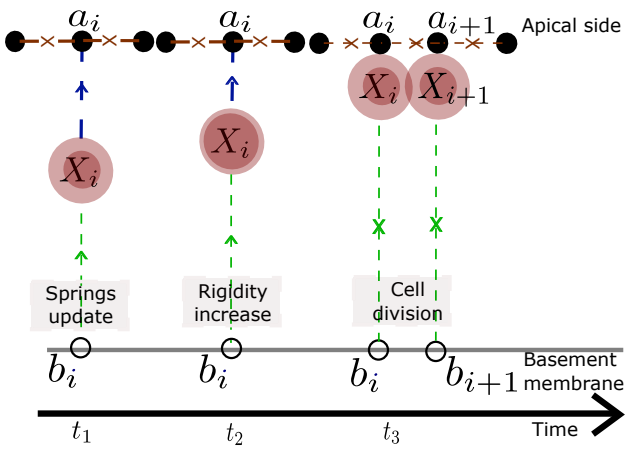


Figure 10: Sketch of the changes occurring to a cell prior to division. The rest length of the springs changes to allow the nucleus to approach the apical side. After a while, the radius of the hard nucleus increases and finally, at division, one cell is substituted by two cells.

Prior to division, the stiffness and rest lengths of the springs of the dividing cell are modified to allow the nucleus to approach the apical side (see Figure 10). After a short time period, the radius of the inner-circle increases, which models an increase in the rigidity of the nucleus. After a while the cell divides, i.e., two cells are introduced in the place of the older cell. Each cell has an internal clock that determines the time for division and its internal state, namely, the nucleus rigidity and the properties of the springs. The position of the nuclei and apical points are subject to Gaussian white noise. The dynamics of the system is driven by noise, cell division, changes in nucleus rigidity and spring characteristics.

To simulate this dynamics, we use the time-stepping scheme developed and tested in Chapter 3. This scheme involves the resolution of a minimization problem with smooth constraints at each time-step. We minimize a potential corresponding to the sum of the energy associated to the springs, the alignment force and the soft nucleus. The minimization is subject to non-overlapping constraints on the nuclei inner-cores, and non-switching and lateral adhesion constraints on the basal points. To solve this minimization problem we use the DAHA algorithm that was developed in Chapter 1. At each time-iteration of the simulation, we perform the following steps:

1. cell division (one cell is replaced by two cells)
2. actualization of internal state of each cell (nucleus rigidity, rest-length and stiffness of springs),
3. adding noise to the position of the nuclei and apical points
4. restoration of the mechanical equilibrium (minimization algorithm).

We then investigate for which region of the parameter space we get a qualitative agreement with the laboratory data acquired by the team of biologists on the neuroepithelium. As much as possible the values of the parameters have been estimated from experimental data.

The model has provided important output about the mechanics underlying the tissue. In particular, it suggests that the presence of noise in the system is fundamental to allow the movement of nuclei in a crowded environment. Another observation is that the speed of actualization of the springs performed in step 2 above seems to control both the time it takes for the nucleus to reach the apical side before division and the height of the tissue. These observations can be compared with experimental results and help in the design of new laboratory experiments.

In the future one should develop more quantitative comparisons between the model and experimental data to calibrate and validate the model. In particular, it would be interesting to explore the potential existence of bifurcations on the model parameters and see if they would relate to *in vivo* observations.

In a longer term the model could be used to explore other biologically relevant situations, such as cancer development. In this case, the model might prove particularly useful, by allowing *in silico* experiments which could be technically challenging *in vivo*. For instance, the effect of finely tuning cell heterogeneity within the tumour or assessing the effect of tumour size on tumour dynamics and on healthy neighbouring cells.

0.4.3 Continuum models for congested systems (Part III)

In Part III we use the intuition developed in Parts I and II to design a continuum approach for congested systems. We assume that the system is at all times at a state of mechanical equilibrium, that corresponds to the solution of a minimization problem. The evolution of the system is then driven by the evolution of an external confining potential and by a change in average particle volume. The question we want to address is what motion results from this situation.

0.4.3.1 A new continuum theory for incompressible swelling materials (Chapter 5)

Chapter 5 is based on a collaboration with Pierre Degond, Sara Merino-Aceituno and Mickaël Nahon. The aims of this Chapter are:

- To develop a continuum model for large and dense particle systems based on the rules introduced in Part II at the particle level.
- To characterize a generic solution.

We develop a continuum model for incompressible swelling materials, which intends to be a first step towards the development of more realistic models of swelling gels or tumours. We start by considering a discrete model in order to motivate the continuum description. Similarly to the models presented in Part II, we consider a system made of N finite-sized particles at equilibrium in a confining potential constrained by the non-overlapping constraints. We then assume that the particle volume and the confining potential may vary with time and that the particles follow this evolution adiabatically by remaining at any time at mechanical equilibrium. We refer to [112] for a discussion of the biological relevance of this description and to Chapter 4 for an example of a discrete model developed within this particular setting.

We now formulate a similar problem at the continuum level. We consider a continuum density $n(x,t)$ for a population of particles having finite average volume $\tau(x,t)$, where $x \in \mathbb{R}^d,\ d = 1,2,3$ represents the position and $t \geq 0$ the time. We assume that the particle density satisfies the total mass constraint

$$\int_{\mathbb{R}^d} n(x,t)dx = N, \tag{22}$$

where $N > 0$ is fixed, the non-negativity constraint

$$n(x,t) \geq 0 \tag{23}$$

and the maximal density constraint

$$n(x,t)\tau(x,t) \leq 1. \tag{24}$$

Moreover, we assume that the system is confined by an external potential $V(x,t,\tau(x,t))$ whose associated mechanical energy is given by

$$F_t[n] = \int_{\mathbb{R}^d} V(x,t,\tau(x,t))\, n(x,t)dx, \tag{25}$$

The state of the system at any given time t corresponds therefore to the solution of the following minimization problem:

$$\begin{aligned} &\text{Find } n(\cdot,t):\ x \in \mathbb{R}^d \mapsto n(x,t) \in \mathbb{R} \ \text{ a solution of:} \\ &\min\big\{F_t[n(\cdot,t)] \quad | \quad n(\cdot,t) \geq 0, \quad n(\cdot,t)\tau \leq 1 \ \text{ and } \ \int_{\mathbb{R}^d} n(x,t)dx = N\big\}, \end{aligned} \tag{26}$$

for $\tau : (x,t) \in \mathbb{R}^d \times [0,\infty) \mapsto \tau(x,t) \in \mathbb{R}_+$ and $N > 0$ given.

We start by characterizing the resulting equilibrium density. We prove that the solution to (26) is unique, contrarily to the discrete case (see section 0.4.1). The formulas we obtain show that the particles fill all the energy level sets of the potential V while keeping the non-overlapping condition saturated (i.e. the density being equal to the packing density)

until they reach a maximal energy value related to the total number of particles. Indeed, the potential V tends to bring all particles towards its points of global minimum. However, the non-overlapping constraint prevents the particles to pile up at these points and forces them to occupy increasingly higher potential values. They do so until the total number of particles has been exhausted. At this point, the medium has reached its outer boundary and it is therefore limited by the level set that encloses a number of particles exactly equal to the total number N of available particles in the system.

Then, we assume that the average particle volume and of the confinement potential depend on time. Assuming that the system moves adiabatically and remains at any time at mechanical equilibrium, we can compute the continuum velocity. More precisely, we determine this velocity by applying two heuristic principles directly connected to the previous non-overlapping heuristics. The first heuristics is that particles cannot swap their positions. Indeed, at the packing state, there is not enough space for two spherical particles to undertake the maneuver required to swap their position. This heuristics provides the component of the velocity *normal* to the level sets of the potential V. At this point, we prove that the medium velocity is consistent with the continuity equation. In order to determine the component of the velocity *tangent* to the level sets of the potential V, we invoke a second heuristics, namely that the sequence of minimization problems over time will favour a continuous particle motion rather than jumps which would generate large velocities. In continuum terms, this means that the velocity should obey an energy minimization principle. We show that this principle determines the tangential velocity in a unique way as the tangential gradient along the potential level sets of a certain velocity potential (not to be confused with the confinement potential). This velocity potential is found by inverting a Laplace-Beltrami operator on each of the level sets.

This approach is different from the classical one relying on Darcy's law. One consequence of the use of Darcy's law for incompressible swelling materials is that, at the medium boundary, the continuum velocity is normal to the boundary [146]. Contrarily, in our framework the velocity at the medium boundary does not have to be (and it is not in general) normal to the boundary, due to the presence of a non-trivial tangential velocity component.

This new modelling approach opens many exciting new research directions, from theory to numerics and modelling to applications. A (non-exhaustive) list of future directions which will be investigated in forthcoming works include the following: adding cell division; considering a potential V that involves a contribution from particle interaction such as attachment between nearby cells; coupling with chemical fields; introducing boundary fuzziness; introducing a statistical description of particle volume sizes leading to a kinetic equation; taking into account multiple particle species; deriving a macroscopic model by coarse-graining; developing a numerical approximation of the system and applying to practical systems.

0.5 Structure of the thesis

In this thesis we propose two modelling frameworks to describe congested particle systems at the discrete (Parts I and II) and continuum (Part III) levels. In the first framework, the system is described at the particle level (microscopic model), while in the second framework the system is described at the number density level (macroscopic model). The two frameworks are based on simple heuristic rules on the interaction between particles that are likely to be obeyed in generic congested systems. New computational and mathematical tools are developed in order to study the static equilibria (Part I and Part III) and evolution (Part II and Part III) of these systems. The static equilibrium corresponds to a solution of a minimization problem. In the discrete framework, such minimization problem is typically non-convex, which brings additional difficulties, as no universal numerical methods exist for these problems and therefore, they need to be treated case by case. Part I of this thesis is exclusively dedicated to the development and study of algorithms to find a numerical solution to this problem.

In Chapter 1 (Part I) we propose a new algorithm, that we call DAHA, for non-convex optimization problems. We develop a numerical study on the convergence and speed of convergence of this algorithm in the context of a sphere packing problem in 2 and 3 spatial dimensions. We develop a comparative study with existing methods belonging to the class of Linearly Constrained Lagrangian.

In Chapter 2 (Part I) we find the analytical solution to the sphere packing problem referred in the previous paragraph in dimension 1, which is unique up to permutations and translations. We then consider a first order formulation of the DAHA and we prove its convergence in dimension 1 towards the theoretical solution. We also explore analytically existing methods, namely, the Uzawa-Arrow-Hurwicz and augmented Lagrangian and compare them with our method.

In Chapter 3 (Part II) we consider the dynamics of N circles undergoing ballistic aggregation in a torus. We study an event-driven (ED) and several time-stepping (TS) algorithms in the case where the density of particles is high. The aim is to test and compare various algorithms to simulate congested systems. We consider TS schemes involving the resolution of a sphere packing problem at each time-iteration, which are obtained by combining smooth and non-smooth form of the constraints and rigid and non-rigid growing aggregates. The sphere packing problem is solved numerically using the DAHA presented in Chapter 1. Numerical experiments show that the ED becomes computationally very expensive as the density and the number of particles increases. The TS scheme with the lowest computational time is the one that uses the smooth form of the constraints and non-rigid clusters. We then select the smooth form of the constraints to be used in the simulation of the dynamics of a cell tissue in Chapter 4.

In Chapter 4 (Part II) we develop a mechanical model of a densely packed epithelial cell tissue in collaboration with a team of biologists led by Eric Theveneau (University Paul

Sabatier, Toulouse), who run *in vivo* lab experiments. The aim is to identify the main mechanical properties that lead to the observed dynamics and shape of the tissue. The algorithm to simulate the model is the TS scheme with smooth form of the constraints that we selected in Chapter 3. The results of the model are in good agreement with experimental data. The model has motivated the formulation of new hypotheses about the shape and evolution of this tissue.

In Chapter 5 (Part III) we propose a continuum description of swelling materials. The aim is to describe the evolution of the particle density in a congested system where the particles are confined by a potential energy. The motion is driven by the evolution of the potential energy and changes in the average particle size. Similarly to the discrete models presented in Chapters 3 and 4, we describe the static equilibrium as a solution to a minimization problem with maximal density constraint. Contrarily to the minimization problem associated to the discrete model presented in Part I, in the continuous model the minimization problem is convex and we are able to compute the only solution analytically. We then compute the medium velocity under two extra assumptions: particles cannot swap positions (non-swapping condition) and they favour a continuous trajectory rather than with jumps (principle of minimal displacement).

A schematic organization of the topics developed in the thesis is presented in Figure 11. The topics are classified in four categories, resulting from the combination of two aspects:

- the type of model: discrete/particle-based or continuous/density-based (represented in the horizontal axis),
- the state of the system: statics/time-independent or dynamics/time-dependent (represented in the vertical axis).

Part I (Chapters 1 and 2) refers to discrete models describing the static equilibrium of congested systems. Part II (Chapters 3 and 4) refers to discrete models describing the dynamics of congested systems. Part III (Chapter 5) refers to continuum models describing both the static equilibrium and dynamics of congested systems.

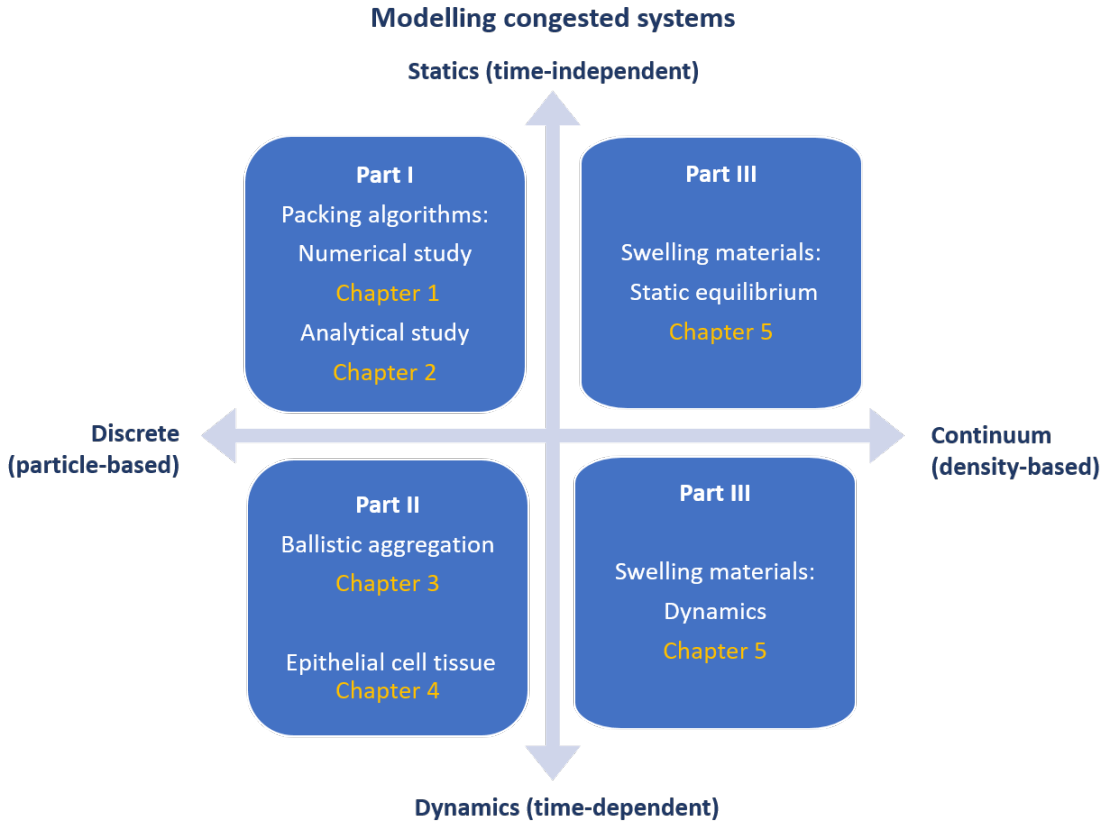


Figure 11: Classification of the Chapters of this thesis and how they relate to each other. The horizontal axis refers to the spatial scale considered in the models: from microscopic (discrete particle-based models, Chapters 1 to 4) to macroscopic (continuum density based models, Chapter 5). The vertical axis refers to the state of the system: static equilibrium (Chapters 1, 2 and 5) versus dynamics (Chapters 3, 4 and 5).

Part I

Algorithms for packing problems

1

Damped Arrow-Hurwicz algorithm for sphere packing

We consider algorithms that, from an arbitrarily sampling of N spheres (possibly overlapping), find a close packed configuration without overlapping. These problems appear in the simulation of particle systems with congestion presented in Part II and they can be formulated as minimization problems with non-convex constraints. For such packing problems, we observe that the classical iterative Arrow-Hurwicz algorithm does not converge. We derive a novel algorithm from a multi-step variant of the Arrow-Hurwicz scheme with damping. We compare this algorithm with classical algorithms belonging to the class of linearly constrained Lagrangian methods and show that it performs better. We provide an analysis of the convergence of these algorithms in the simple case of two spheres in one spatial dimension. Finally, we investigate the behaviour of our algorithm when the number of spheres is large in two and three spatial dimensions. This algorithm is further analysed in Chapter 2.

The results presented in this Chapter have been published[*] *in Journal of Computational physics [45] in collaboration with Sebastien Motsch*[†] *and Pierre Degond*[‡].

[†]School of Mathematical & Statistical Sciences, Arizona State University, Tempe, USA
[‡]Department of Mathematics, Imperial College London, UK

1.1 Introduction

Particle packing problems can be encountered in many different systems, from the formation of planets or cells in live tissues to the dynamics of crowds of people. In particular, they have been widely investigated in the study of granular media [117], glasses [181] and liquids [85]. More recently, particle packings have also been used in biology, in the study of bacterial colonies, [107] and social sciences, in the estimation of a maximum safe packing in a crowd [39].

Packing problems lead to NP-hard non-convex optimization problems [89], due to the non-convexity of the exclusion volume constraints. In the literature, we find numerical studies involving particles with various shapes such as ellipses [180] or even non-convex particles [60]. Depending on the degree of symmetry of the particles shape and the packing domain, the optimal solution may not be unique, since permutations, rotations or reflections may generate equivalent solutions. An example where all these symmetries occur is the packing of a finite number of spheres in the Euclidean space, which corresponds to the situation we explore here. Moreover, there are local minima and their number increases as the number of particles and the spatial dimensions increases.

We consider algorithms that, given an initial configuration of N spheres (possibly overlapping), find a nearby packed configuration without overlapping. This solution is not in general a global optimum, which is an important feature regarding the applications we are interested in. Indeed, in many natural systems individuals or particles only seek to achieve a locally optimal solution. Therefore, it is more likely that they reach a local configuration that does not necessarily correspond to a global optimum. By combining our method with, for example, simulated annealing techniques [1], we could convert our algorithms into global minimum search algorithms. It is however not the objective we pursue here.

Classical procedures for non-convex optimization include Uzawa-Arrow-Hurwicz type algorithms [17], augmented Lagrangian [23, 25], linearly constrained Lagrangian (LCL), sequential quadratic programming (SQP) [134], among others. For the type of problems we consider here, the Hessian matrix of the function to be minimized is not always positive definite (see the example presented in section 1.3). Consequently, the SQP and Uzawa-Arrow-Hurwicz algorithms do not always converge. On the other hand, the LCL methods converge very quickly if the number of particles is small. However we are interested in the case where this number becomes large.

In [53, 86] the authors study the shape of three dimensional clusters of atoms under the effect of soft potentials by using molecular dynamics. This approach differs from ours with regard to the non-overlapping constraints, which are approximated by soft potentials, producing soft dynamics. Although being more costly when dealing with a large number of particles, we have opted by the hard dynamics approach, since it allows for a higher precision in the treatment of the constraints. This proves effective when

dealing with interaction between rigid bodies, where the effect of the rigid boundaries plays an important role. This motivates the present work.

We start in section 1.2.1 by presenting two formulations of the problem. The first one is the classical minimization approach. The second one considers a constrained dynamical system in the spirit of [122]. We also present two equivalent types of non-overlapping constraints involving smooth or non-smooth functions which are found in the literature [89, 4]. To solve the non-convex minimization problems arising in these formulations, we first consider in section 1.2.3 the classical Arrow-Hurwicz algorithm (AHA). In section 1.2.4 we introduce a novel multi-step scheme based on a second-order ODE interpretation of the minimization problem: the damped Arrow-Hurwicz algorithm (DAHA). We test the DAHA against two methods taken from the widely known class of linearly constrained Lagrangian algorithms [134, 122]. These algorithms consist of a sequence of convex minimization problems, for we refer to them as *nested algorithms (NA)* and they shall be referred to as the NAP and NAV. The convergence of the four algorithms (the AHA, DAHA, NAP and NAV) is analyzed in section 1.3 for the case of two spheres in one dimension. In section 1.4 the algorithms are numerically compared for the cases of many spheres in two dimensions. A brief numerical study of the packing density in two and three dimensions is also presented. Finally, conclusions and future works are presented in section 1.5. We also refer to Chapter 2 for a more detailed analysis of the minimization problem. In particular we prove that minimizers are not saddle points of the Lagrangian. In the present Chapter we investigate the convergence of the algorithms numerically.

1.2 The damped Arrow-Hurwicz algorithm (DAHA)

1.2.1 Minimization problems for sphere packing

We first recall two different formulations of generic minimization problems. Let N and b be two given positive integers. We consider first the problem of finding a configuration $\bar{\mathbf{X}}$ such that

$$\bar{\mathbf{X}} \in \underset{\phi_{k\ell}(\mathbf{X})\leq 0,\ k,\ell=1,...,N,\ k<\ell}{\operatorname{argmin}} W(\mathbf{X}). \tag{1.1}$$

where $W : \mathbb{R}^{bN} \to \mathbb{R}$ is a convex function (not necessarily strictly convex). The functions $\phi_{k\ell} : \mathbb{R}^{bN} \to \mathbb{R},\ k, \ell = 1, ..., N,\ k < \ell$ are continuous but not necessarily convex. We suppose that W has an attainable minimum in the set of admissible solutions $\{\mathbf{X} \in \mathbb{R}^{bN} | \phi_{k\ell}(\mathbf{X}) \leq 0,\ k, \ell = 1, ..., N,\ k < \ell\}$. In these conditions, $\bar{\mathbf{X}}$ exists but may not be unique. We also assume that $\phi_{k\ell},\ k, \ell = 1, ..., N,\ k < \ell$ and W are C^1 functions in the neighbourhood of $\bar{\mathbf{X}}$. In what follows, d will denote the diameter of a sphere, N the number of spheres, b the spatial dimension, $\mathbf{X}$ the position of the center of the spheres and $\phi_{k\ell}$ the non-overlapping constraint functions between the kth and ℓth spheres. The non-

overlapping constraints for a system of identical spheres in $\mathbb{R}^b$ can be expressed by means of a smooth or a non-smooth function as specified bellow. Although leading to equivalent constraints, each form has an impact on the convergence of the numerical method towards a local minimizer, as we will see in sections 1.3 and 1.4.

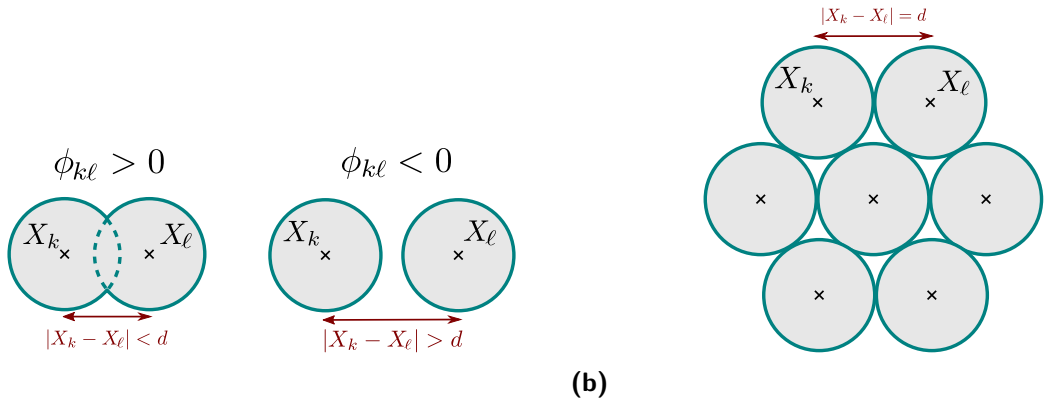
Definition 7. *We call* ***non-smooth form of the constraint functions (NS)*** *the following function*

$$\phi_{k\ell}(\mathbf{X}) = d - |X_k - X_\ell|, k, \ell = 1, ..., N, \ k \neq \ell$$

and ***smooth form of the constraint functions (S)*** *the following function*

$$\phi_{k\ell}(\mathbf{X}) = d^2 - |X_k - X_\ell|^2, k, \ell = 1, ..., N, \ k \neq \ell.$$

An illustration of the non-overlapping constraints, as well as, a possible solution for $N = 7$ in dimension $b = 2$ are presented in Figure 1.1.



(a) **(b)**

Figure 1.1: Representation of the non-overlapping constraints, 1.1a, and a possible optimal solution of (1.1) for $N = 7$, 1.1b.

We now present a second formulation consisting in solving a minimization problem associated with a discrete dynamical system which has $\bar{\mathbf{X}}$ as a fixed point. Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^b$. The problem is formulated iteratively: given an initial configuration $\mathbf{X}^0 = \{X_i^0\}_{i=1,...,N}$, we pass from iterate $\mathbf{X}^p$ to iterate $\mathbf{X}^{p+1}$ as follows

$$\begin{cases} \mathbf{X}^{p+1} = \mathbf{X}^p + \tau \ \mathbf{V}^{p+1} & (1.2) \\ \mathbf{V}^{p+1} \in \underset{\phi_{k\ell}(\mathbf{X}^p + \tau \ \mathbf{V}) \leq 0, \ k,\ell=1,...,N, \ k<\ell}{\operatorname{argmin}} \dfrac{1}{2} \displaystyle\sum_{i=1}^{N} |V_i + \nabla_{X_i} W(\mathbf{X}^p)|^2, & (1.3) \end{cases}$$

where $\tau > 0$ is a given parameter and $\mathbf{V} = \{V_i\}_{i=1,...,N}$. At each iteration $p + 1$, we look for the fictitious velocity $\mathbf{V}^{p+1}$ that is close to the steepest descent $-\nabla_{X_i} W(\mathbf{X}^p)$ and such that the iterate $\mathbf{X}^{p+1} = \mathbf{X}^p + \tau \mathbf{V}^{p+1}$ is a non-overlapping configuration. The algorithm

stops when the velocity reaches zero, yielding a fixed point $\widetilde{\mathbf{X}}$ satisfying

$$0 \in \underset{\phi_{k\ell}(\widetilde{\mathbf{X}}+\tau\ \mathbf{V})\leq 0,\ k,\ell=1,\dots,N,\ k<\ell}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{N} |V_i + \nabla_{X_i} W(\widetilde{\mathbf{X}})|^2. \tag{1.4}$$

In particular $\widetilde{\mathbf{X}}$ satisfies the non-overlapping constraint, $\phi_{k\ell}(\widetilde{\mathbf{X}}) \leq 0$. Note that the set defined by the constraints is compact, so the minima of (1.4) exist, but may not be unique. As we will see in the next Section, the Lagrangian formulation of the two problems (1.1) and (1.4) are equivalent.

1.2.2 Lagrangian formulation

The minimization problem (1.1) can be formulated in terms of the Lagrangian $\mathcal{L} : \mathbb{R}^{bN} \times (\mathbb{R}_0^+)^{N(N-1)/2} \to \mathbb{R}$ defined by

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}) = W(\mathbf{X}) + \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}\phi_{k\ell}(\mathbf{X}),$$

where $\boldsymbol{\lambda} = \{\lambda_{k\ell}\}_{k,\ell=1,\dots,N,\ k<\ell}$ represents the set of Lagrange multipliers.

We start by proving that the Abadie constraint qualification (ACQ) introduced in section 0.3.2.1 holds at $\bar{\mathbf{X}}$ for the smooth and non-smooth constraints $\phi_{k\ell}$ defined in definition 7.

Proposition 8. *Let $\bar{\mathbf{X}} = \{\bar{X}_i\}_{i=1,\dots,N}$ be a solution to the problem* (1.1)*. Then the ACQ holds at $\bar{\mathbf{X}}$.*

Proof. Consider the tangent and the linearized cones $T(\bar{\mathbf{X}})$ and $L(\bar{\mathbf{X}})$ defined in definition 2 in section 0.3.2.1. We already know that $T(\bar{\mathbf{X}}) \subseteq L(\bar{\mathbf{X}})$. We now prove the other inclusion for each form of the constraints.

- Smooth form of the constraint functions:

Let k, ℓ be such that $\phi_{k\ell}(\bar{\mathbf{X}}) = 0$ and let " $\cdot$ " and "$\langle\cdot\rangle$" denote the inner-product in $\mathbb{R}^b$ and $\mathbb{R}^{bN}$, respectively. For any $\mathbf{D} = \{D_i\}_{i=1,\dots,N} \in L(\bar{\mathbf{X}}) \subset \mathbb{R}^{bN}$, we have

$$\begin{aligned} 0 &\geq \langle \nabla\phi_{k\ell}(\bar{\mathbf{X}}), \mathbf{D}\rangle \\ &= -2(\bar{X}_k - \bar{X}_\ell)\cdot D_k - 2(\bar{X}_\ell - \bar{X}_k)\cdot D_\ell \\ &= -2(\bar{X}_k - \bar{X}_\ell)\cdot (D_k - D_\ell), \end{aligned}$$

which implies that $(\bar{X}_k - \bar{X}_\ell) \cdot (D_k - D_\ell) \geq 0$. Consequently, for any $t^n \geq 0$,

$$\begin{aligned}
\phi_{k\ell}(\bar{\mathbf{X}} + t^n \mathbf{D}) &= d^2 - |\bar{X}_k + t^n D_k - \bar{X}_\ell - t^n D_\ell|^2 \\
&= d^2 - (\bar{X}_k - \bar{X}_\ell + t^n(D_k - D_\ell)) \cdot (\bar{X}_k - \bar{X}_\ell + t^n(D_k - D_\ell)) \\
&= d^2 - |\bar{X}_k - \bar{X}_\ell|^2 - 2t^n(\bar{X}_k - \bar{X}_\ell) \cdot (D_k - D_\ell) - (t^n)^2 |D_k - D_\ell|^2 \\
&= -t^n(\bar{X}_k - \bar{X}_\ell) \cdot (D_k - D_\ell) - (t^n)^2 |D_k - D_\ell|^2 \qquad (1.5) \\
&\leq 0
\end{aligned}$$

where the step (1.5) comes from $\phi_{k\ell}(\bar{\mathbf{X}}) = 0$. Now, let k, ℓ be such that $\phi_{k\ell}(\bar{\mathbf{X}}) < 0$. As $\phi_{k\ell}$ is a C^1 function in a neighbourhood of $\bar{\mathbf{X}}$, we have that $\phi_{k\ell}(\bar{\mathbf{X}} + \mathbf{V}) \leq 0$, for any $\bar{\mathbf{X}} + \mathbf{V}$ sufficiently close to $\bar{\mathbf{X}}$. Therefore, there exist sequences $\{\mathbf{D}^n\}_n \to \mathbf{D}$ and $\{t^n\}_n \to 0^+$ such that $\phi_{k\ell}(\bar{\mathbf{X}} + t^n \mathbf{D}^n) \leq 0$, for all n and $k, \ell = 1, ..., N,\ k < \ell$. We conclude that $\mathbf{D} \in T(\bar{\mathbf{X}})$.

- Non-smooth form of the constraint functions:

Using a similar reasoning, we conclude that $(\bar{X}_k - \bar{X}_\ell) \cdot (D_k - D_\ell) \geq 0$. Consequently, for any $t^n \geq 0$,

$$\begin{aligned}
\phi_{k\ell}(\bar{\mathbf{X}} + t^n \mathbf{D}) &= d - |\bar{X}_k + t^n D_k - \bar{X}_\ell - t^n D_\ell| \\
&= d - [(\bar{X}_k - \bar{X}_\ell + t^n(D_k - D_\ell)) \cdot (\bar{X}_k - \bar{X}_\ell + t^n(D_k - D_\ell))]^{1/2} \\
&= d - [|\bar{X}_k - \bar{X}_\ell|^2 + 2t^n(\bar{X}_k - \bar{X}_\ell) \cdot (D_k - D_\ell) + (t^n)^2 |D_k - D_\ell|^2]^{1/2} \\
&= d - [d^2 + 2t^n(\bar{X}_k - \bar{X}_\ell) \cdot (D_k - D_\ell) + (t^n)^2 |D_k - D_\ell|^2]^{1/2} \\
&\leq 0.
\end{aligned}$$

The proof follows now exactly the same path as in the smooth case. ∎

By theorem 4 from section 0.3.2.1, there exists $\bar{\boldsymbol{\lambda}} \in (\mathbb{R}_0^+)^{N(N-1)/2}$ such that $(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}})$ is a critical-point of the Lagrangian, namely, $(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}})$ satisfies the *KKT-conditions* [148, 101]:

$$\begin{cases}
\nabla_{X_i} \mathcal{L}(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}}) = 0,\ i = 1, ..., N \\
\left(\nabla_{\lambda_{k\ell}} \mathcal{L}(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}}) = 0 \ \text{ and } \bar{\lambda}_{k\ell} \geq 0\right) \text{ or } \left(\nabla_{\lambda_{k\ell}} \mathcal{L}(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}}) < 0 \text{ and } \bar{\lambda}_{k\ell} = 0\right), \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k, \ell = 1, ..., N, k < \ell
\end{cases}$$

which is equivalent to

$$\begin{cases}
\nabla_{X_i} W(\bar{\mathbf{X}}) + \sum_{k,\ell \in \{1,...,N\},\ k<\ell} \bar{\lambda}_{k\ell} \nabla_{X_i} \phi_{k\ell}(\bar{\mathbf{X}}) = 0,\ i = 1, ..., N \\
\left(\phi_{k\ell}(\bar{\mathbf{X}}) = 0 \text{ and } \bar{\lambda}_{k\ell} \geq 0\right) \text{ or } \left(\phi_{k\ell}(\bar{\mathbf{X}}) < 0 \text{ and } \bar{\lambda}_{k\ell} = 0\right), \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k, \ell = 1, ..., N, k < \ell.
\end{cases} \qquad (1.6)$$

We have reduced our original problem (1.1) to a critical-point system, with a possible enlargement of the set of solutions. In contrast to convex optimization, in the case of

packing problems, these critical-points may not be saddle-points. In Chapter 2 we provide a detailed analysis of this point.

We also formulate the minimization problem (1.3) in terms of a Lagrangian $\mathcal{L}^p$,

$$\mathcal{L}^p(\mathbf{V}, \boldsymbol{\mu}) = \frac{1}{2}\sum_{i=1}^{N} |V_i + \nabla_{X_i} W(\mathbf{X}^p)|^2 + \sum_{k,\ell\in\{1,...,N\},\ k<\ell} \mu_{k\ell}\phi_{k\ell}(\mathbf{X}^p + \tau\ \mathbf{V}),$$

where $\boldsymbol{\mu} = \{\mu_{k\ell}\}_{k,\ell=1,...,N,\ k<\ell}$ is the set of Lagrange multipliers associated to the constraints. The gradients of the Lagrangian are given by

$$\begin{aligned}
\nabla_{V_i}\mathcal{L}^p(\mathbf{V}, \boldsymbol{\mu}) &= V_i + \nabla_{X_i} W(\mathbf{X}^p) \\
&\quad + \tau \sum_{k,\ell\in\{1,...,N\},\ k<\ell} \mu_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X}^p + \tau\ \mathbf{V}),\ i = 1, ..., N \\
\nabla_{\mu_{k\ell}}\mathcal{L}^p(\mathbf{V}, \boldsymbol{\mu}) &= \phi_{k\ell}(\mathbf{X}^p + \tau\ \mathbf{V}),\ k, \ell = 1, ..., N.
\end{aligned}$$

The dynamical system is written: $\widetilde{\mathbf{X}}^{p+1} = \widetilde{\mathbf{X}}^p + \tau\ \tilde{\mathbf{V}}^{p+1}$ such that $(\tilde{\mathbf{V}}^{p+1}, \tilde{\boldsymbol{\mu}}^{p+1})$ is a solution of the critical-point problem

$$\begin{cases}
\tilde{V}_i^{p+1} + \nabla_{X_i} W(\widetilde{\mathbf{X}}^p) + \tau \sum\limits_{k,\ell\in\{1,...,N\},\ k<\ell} \tilde{\mu}_{k\ell}^{p+1}\nabla_{X_i}\phi_{k\ell}(\widetilde{\mathbf{X}}^p + \tau\ \tilde{\mathbf{V}}^{p+1}) = 0, \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, ..., N \\
\left(\phi_{k\ell}(\widetilde{\mathbf{X}}^p + \tau\ \tilde{\mathbf{V}}^{p+1}) = 0 \text{ and } \tilde{\mu}_{k\ell}^{p+1} \geq 0\right) \text{ or} \\
\qquad \left(\phi_{k\ell}(\widetilde{\mathbf{X}}^p + \tau\ \tilde{\mathbf{V}}^{p+1}) < 0 \text{ and } \tilde{\mu}_{k\ell}^{p+1} = 0\right),\ k, \ell = 1, ..., N,\ k < \ell
\end{cases} \tag{1.7}$$

Likewise, the fixed point $\widetilde{\mathbf{X}}$ of the dynamical system is defined such that there exists $\tilde{\boldsymbol{\mu}}$ such that

$$\begin{cases}
\nabla_{X_i} W(\widetilde{\mathbf{X}}) + \tau \sum\limits_{k,\ell\in\{1,...,N\},\ k<\ell} \tilde{\mu}_{k\ell}\nabla_{X_i}\phi_{k\ell}(\widetilde{\mathbf{X}}) = 0,\ i = 1, ..., N \\
\left(\phi_{k\ell}(\widetilde{\mathbf{X}}) = 0 \text{ and } \tilde{\mu}_{k\ell} \geq 0\right) \text{ or } \left(\phi_{k\ell}(\widetilde{\mathbf{X}}) < 0 \text{ and } \tilde{\mu}_{k\ell} = 0\right), \\
\qquad\qquad k, \ell = 1, ..., N,\ k < \ell.
\end{cases} \tag{1.8}$$

Then, it is clear that problems (1.6) and (1.8) are equivalent for all values of $\tau > 0$ by setting $\bar{\boldsymbol{\lambda}} = \tau\tilde{\boldsymbol{\mu}}$. However, the choice of τ is important to ensure convergence of the dynamical system (1.7) to the fixed point.

As it will be obvious below, all functions W and $\phi_{k\ell}$ used throughout the Chapter will satisfy the conditions considered in this section. The nonlinear systems (1.6) or (1.7) will have to be solved by an iterative algorithm. We now present the algorithms considered in the Chapter.

1.2.3 The Arrow-Hurwicz algorithm (AHA)

The classical Arrow-Hurwicz iterative algorithm [17] searches a saddle-point of the Lagrangian by alternating steps in the direction of $-\nabla_{\mathbf{X}}\mathcal{L}$ and $+\nabla_{\boldsymbol{\lambda}}\mathcal{L}$. Using this idea, a saddle-point is then a steady-state solution of the Arrow-Hurwicz system of ODE's (AHS) which is defined next.

Definition 9. *The **Arrow-Hurwicz system (AHS)** is defined by*

$$\begin{cases} \dot{X}_i &= -\alpha \left(\nabla_{X_i} W(\mathbf{X}) + \sum\limits_{k,\ell \in \{1,\dots,N\},\ k<\ell} \lambda_{k\ell} \nabla_{X_i} \phi_{k\ell}(\mathbf{X}) \right), \\ & \qquad\qquad\qquad\qquad\qquad\qquad i = 1, \dots, N \qquad (1.9) \\ \dot{\lambda}_{k\ell} &= \begin{cases} 0, & \text{if } \lambda_{k\ell} = 0 \text{ and } \phi_{k\ell}(\mathbf{X}) < 0 \\ \beta \phi_{k\ell}(\mathbf{X}), & \text{otherwise} \end{cases}, \\ & \qquad\qquad\qquad\qquad\qquad\qquad k, \ell = 1, \dots, N,\ k < \ell, \qquad (1.10) \end{cases}$$

*where α and β are positive constants. Considering a small time-step Δt, a semi-implicit Euler discretization scheme of the previous system leads to the **Arrow-Hurwicz algorithm (AHA)**, which is defined iteratively by*

$$\begin{cases} X_i^{n+1} &= X_i^n - \alpha \left[\nabla_{X_i} W(\mathbf{X}^n) + \sum\limits_{k,\ell \in \{1,\dots,N\},\ k<\ell} \lambda_{k\ell}^n \nabla_{X_i} \phi_{k\ell}(\mathbf{X}^n) \right], \\ & \qquad\qquad\qquad\qquad\qquad\qquad i = 1, \dots, N \\ \lambda_{k\ell}^{n+1} &= \max\{0, \lambda_{k\ell}^n + \beta \phi_{k\ell}(\mathbf{X}^{n+1})\},\ k, \ell = 1, \dots, N,\ k < \ell \end{cases} \qquad (1.11)$$

where α and β now correspond to $\tilde{\alpha} = \alpha \Delta t$ and $\tilde{\beta} = \beta \Delta t$ and the tildes have been dropped for simplicity.

The original AHA was formulated using a fully explicit Euler scheme, but it has proved more accurate to use a semi-implicit scheme. Finding a local steady-sate solution of (1.9)-(1.10) in the case of a packing problem has revealed not to be always possible because it often happens that no critical-point is a saddle-point (see Chapter 2, Section 2.4.1). This manifests itself by the existence of periodic solutions of the AHS which do not converge to the critical-point. In order to overcome this difficulty we propose the damped Arrow-Hurwicz algorithm which is presented next. This method is based on a modification of the dynamics of the AHS that transforms an unstable critical-point into an asymptotically stable one. The performance of our method will be tested by comparing with previous approaches [134, 122], which are based on a modification of the Lagrangian by linearly approximating the constraints. These approaches are presented in section 1.2.5.

1.2.4 The damped Arrow-Hurwicz algorithm

In order to avoid periodic solutions we will add a damping term as described below. Note that we are not interested on the transient dynamics of the system, but rather on its asymptotic behaviour.

We propose the following definition.

Definition 10. *We define the* ***damped Arrow-Hurwicz system (DAHS)*** *as*

$$\begin{cases} \ddot{X}_i &= -\alpha^2[\nabla_{X_i} W(\mathbf{X}) + \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X})] \\ &\quad -\alpha\beta \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X})\lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X}) - c\dot{X}_i\ , \\ &\qquad\qquad i = 1,\dots,N \qquad (1.12) \\ \dot{\lambda}_{k\ell} &= \begin{cases} 0, & \text{if } \lambda_{k\ell} = 0 \text{ and } \phi_{k\ell}(\mathbf{X}) < 0 \\ \beta\phi_{k\ell}(\mathbf{X}), & \text{otherwise} \end{cases}, \\ &\qquad\qquad k,\ell = 1,\dots,N,\ k<\ell \qquad (1.13) \end{cases}$$

where α, β and c are positive constants and the ***damped Arrow-Hurwicz algorithm (DAHA)*** *as the corresponding semi-implicit discrete scheme:*

$$\begin{cases} X_i^{n+1} &= \frac{1}{1+c/2}\left(2X_i^n - (1-c/2)X_i^{n-1}\right) \\ &\quad -\frac{\alpha^2}{1+c/2}[\nabla_{X_i} W(\mathbf{X}^n) + \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}^n\nabla_{X_i}\phi_{k\ell}(\mathbf{X}^n)] \\ &\quad -\frac{\alpha\beta}{1+c/2} \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X}^n)\lambda_{k\ell}^n\nabla_{X_i}\phi_{k\ell}(\mathbf{X}^n), \\ &\qquad\qquad i = 1,\dots,N \qquad (1.14) \\ \lambda_{k\ell}^{n+1} &= \max\{0, \lambda_{k\ell}^n + \beta\phi_{k\ell}(\mathbf{X}^{n+1})\},\ k,\ell = 1,\dots,N,\ k<\ell, \qquad (1.15) \end{cases}$$

where α, β and c correspond now to numerical parameters.

Note that the DAHA is a multi-step scheme, since not only one, but two previous configurations $\mathbf{X}^{n-1}$ and $\mathbf{X}^n$ are used to obtain $\mathbf{X}^{n+1}$. By setting $c = 2$, the method is reduced to a one-step method. As observed numerically in Section 1.4, the value $c = 2$ seems to yield the fastest convergence of the algorithm for the particular potential W defined in (1.25). This motivates the analytical study of the corresponding first-order ODE system presented in Chapter 2. For generic potentials it remains though an open question whether this reduction also leads to a faster convergence or not.

A second-order ODE system with damping has previously been proposed within the scope of convex programming [9, 10]. Besides comprising the non-convex case, our approach differs from this with regard to the extra term $\alpha\beta$ in equation (1.12).

In the following we present the derivation of the DAHS. We start by considering the AHS (1.9)-(1.10) presented in the previous section. We then take the second-order version

of (1.9). For each $i = 1, ..., N$ we have

$$\begin{aligned}\ddot{X}_i = \quad & -\alpha \sum_{m=1}^{N} \nabla_{X_m} \big[\nabla_{X_i} W(\mathbf{X}) + \sum_{k,\ell \in \{1,...,N\},\ k<\ell} \lambda_{k\ell} \nabla_{X_i} \phi_{k\ell}(\mathbf{X}) \big] \dot{X}_m \\ & -\alpha \sum_{k,\ell \in \{1,...,N\},\ k<\ell} \dot{\lambda}_{k\ell} \nabla_{X_i} \phi_{k\ell}(\mathbf{X}). \end{aligned} \tag{1.16}$$

Using (1.10), we can replace $\dot{\lambda}_{k\ell}$ in (1.16) by $\beta \phi_{k\ell}(\mathbf{X}) H(\lambda_{k\ell})$, where H is the Heaviside function. Moreover, in order to keep the same steady states as the AHS, we replace $H(\lambda_{k\ell})$ by $\lambda_{k\ell}$, as at equilibrium $\lambda_{k\ell}\phi_{k\ell} = 0$ holds. Note that other choices could have been made, such as a power of $\lambda_{k\ell}$ for instance, which would have influence in the speed of convergence of the algorithm, however we do not explore this aspect further here. We get

$$\ddot{X}_i = -\alpha \sum_{m=1}^{N} \nabla_{X_m} \big[\nabla_{X_i} W(X) + \sum_{k,\ell \in \{1,...,N\},\ k<\ell} \lambda_{k\ell} \nabla_{X_i} \phi_{k\ell}(\mathbf{X}) \big] \dot{X}_m \tag{1.17}$$

$$-\alpha\beta \sum_{k,\ell \in \{1,...,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X}) \lambda_{k\ell} \nabla_{X_i} \phi_{k\ell}(\mathbf{X}). \tag{1.18}$$

It turns out that passing to the second-order introduces exponentially growing modes (see Remark 1.2.1).

Remark 1.2.1. *Consider the simple ODE $\dot{u} = -\alpha u$ whose solution is $u(t) = u_0 e^{-\alpha t}$, where u_0 is the initial configuration. Differentiating both sides of the equation and substituting $\dot{u}$ by $-\alpha u$ yields $\ddot{u} = \alpha^2 u$, whose solution includes now an exponentially growing mode: $u(t) = c_1 e^{-\alpha t} + c_2 e^{\alpha t}$, where c_1 and c_2 are real constants determined by the initial configurations.*

In order to remove these modes, we replace the term in (1.17) by a simple second-order dynamics in the force field given by the right hand side of (1.9). We get:

$$\begin{aligned}\ddot{X}_i = \quad & -\alpha^2 \left[\nabla_{X_i} W(\mathbf{X}) + \sum_{k,\ell \in \{1,...,N\},\ k<\ell} \lambda_{k\ell} \nabla_{X_i} \phi_{k\ell}(\mathbf{X}) \right] \\ & -\alpha\beta \sum_{k,\ell \in \{1,...,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X}) \lambda_{k\ell} \nabla_{X_i} \phi_{k\ell}(\mathbf{X}). \end{aligned} \tag{1.19}$$

In this way, the first term is proportional to $-\nabla_{\mathbf{X}} \mathcal{L}$, which avoids divergent modes and stabilizes the dynamics. Now, we just add a velocity damping term in the form of $-c\dot{X}_i$ and we finally obtain (1.12). We end up with the system (1.12)-(1.13).

Remark 1.2.2. *We can interpret equation* (1.12) *as a second-order dynamics version of* (1.9)*. Denoting by T_1 and T_2 the terms in* (1.19) *which are multiplied by $-\alpha^2$ and $-\alpha\beta$,*

respectively, we recover (1.9) *in an over-damped limit* $\epsilon\left[\ddot{X}_i + \alpha\beta T_2\right] = -\alpha^2 T_1 - c\dot{X}_i$, *with* $\epsilon \to 0$ *and* $c = 1$.

Proposition 11. *The AHS* (1.9)-(1.10) *and the DAHS* (1.12)-(1.13) *have the same equilibrium solutions.*

Proof. If $(\boldsymbol{\lambda}^*, \mathbf{X}^*)$ is a steady state of the AHS, then either $\phi_{k\ell}(\mathbf{X}^*) = 0$ or $\lambda^*_{k\ell} = 0$. Consequently, $\lambda_{k\ell}\phi_{k\ell}(\mathbf{X}^*) = 0$, which implies that the second part of equation (1.12) is null and $\ddot{\mathbf{X}}^* = 0$. Using a similar argument we conclude that a steady state of DAHS is also a steady state of AHS. ∎

1.2.5 Previous approaches

A common approach to solve the generic minimization problems (1.1) and (1.2)-(1.3) is based on the linearization of the constraint functions $\phi_{k\ell}$ around a certain configuration $\mathbf{X}^p$, which we denote by $\phi^p_{k\ell}(\mathbf{X})$, i.e,

$$\phi^p_{k\ell}(\mathbf{X}) = \phi_{k\ell}(\mathbf{X}^p) + \nabla_{\mathbf{X}}\phi_{k\ell}(\mathbf{X}^p)\cdot(\mathbf{X} - \mathbf{X}^p). \tag{1.20}$$

The solution $\mathbf{X}^{p+1}$ of the resulting linearly constrained optimization problem is used to improve the linearization of the constraint functions and this process is iterated until convergence. Note that this transformation turns the non-convex minimization problems (1.1) and (1.2)-(1.3) into a sequence of convex problems, for which there are many tools available [24]. We have chosen the Arrow-Hurwicz algorithm, however, any other method for convex optimization problems would suit our purpose.

This method belongs to the class of linearly constrained Lagrangian (LCL) methods [134] which have been used for large constrained optimization problems.

1.2.5.1 The nested algorithm for the positions (NAP)

Consider the system (1.6) with linearized constraint functions. We propose the following definition.

Definition 12 (**Nested Algorithm for the Positions (NAP)**). *Let* $(\mathbf{X}^p, \boldsymbol{\lambda}^p)$ *be given. Define* $\mathbf{X}^{p,0} = \mathbf{X}^p$, $\boldsymbol{\lambda}^{p,0} = \boldsymbol{\lambda}^p$ *and* $\phi^p_{k\ell}$ *as in* (1.20). *For a given* $(\mathbf{X}^{p,n}, \boldsymbol{\lambda}^{p,n})$, *let the step of the inner-loop be defined as*

$$\begin{cases} X_i^{p,n+1} & = X_i^{p,n} - \alpha\left[\nabla_{X_i}W(\mathbf{X}^{p,n}) + \sum\limits_{k,\ell\in\{1,\dots,N\},\ k<\ell}\lambda^{p,n}_{k\ell}\nabla_{X_i}\phi^p_{k\ell}(\mathbf{X}^{p,n})\right], \\ & \qquad\qquad i = 1,\dots,N \qquad (1.21) \\ \lambda^{p,n+1}_{k\ell} & = \max\left\{0, \lambda^{p,n}_{k\ell} + \beta\phi^p_{k\ell}(\mathbf{X}^{p,n+1})\right\},\ k,\ell = 1,\dots,N,\ k<\ell, \qquad (1.22) \end{cases}$$

then $(\mathbf{X}^{p+1}, \boldsymbol{\lambda}^{p+1}) = \lim_{n\to\infty}(\mathbf{X}^{p,n}, \boldsymbol{\lambda}^{p,n})$.

If we only compute one step of the inner-loop per iteration of the outer-loop we get a variant of the AHA formulation, where $\phi_{k\ell}(\mathbf{X}^{p+1})$ is replaced by $\phi^p_{k\ell}(\mathbf{X}^{p+1})$ in (1.22).

1.2.5.2 The nested algorithm for the velocities (NAV)

We consider the minimization problem (1.7) with linearized constraint functions.

Definition 13 (Nested Algorithm for the Velocities (NAV)). *Let $\tau > 0$ and $(\mathbf{X}^p, \mathbf{V}^p, \boldsymbol{\mu}^p)$ be given. Define $\mathbf{V}^{p,0} = \mathbf{V}^p$, $\boldsymbol{\mu}^{p,0} = \boldsymbol{\mu}^p$ and $\phi^p_{k\ell}$ as in (1.20). For a given $(\mathbf{V}^{p,n}, \boldsymbol{\mu}^{p,n})$, let the step of the inner-loop be defined as*

$$\begin{cases} V_i^{p,n+1} = V_i^{p,n} \\ -\alpha \left(V_i^{p,n} + \nabla_{X_i} W(\mathbf{X}^p) + \tau \sum\limits_{k,\ell \in \{1,\dots,N\},\ k<\ell} \boldsymbol{\mu}^{p,n}_{k\ell} \nabla_{X_i} \phi^p_{k\ell}(\mathbf{X}^p + \tau \mathbf{V}^{p,n}) \right), \\ \qquad\qquad i = 1, ..., N \qquad (1.23) \\ \mu^{p,n+1}_{k\ell} = \max\left\{0, \mu^{p,n}_{k\ell} + \beta \phi^p_{k\ell}(\mathbf{X}^p + \tau \mathbf{V}^{p,n+1})\right\}, k, \ell = 1, ..., N, \qquad (1.24) \end{cases}$$

then $(\mathbf{V}^{p+1}, \boldsymbol{\mu}^{p+1}) = \lim_{n\to\infty}(\mathbf{V}^{p,n}, \boldsymbol{\mu}^{p,n})$ and $\mathbf{X}^{p+1} = \mathbf{X}^p + \tau\ \mathbf{V}^{p+1}$.

The NAV corresponds to an adaptation of the method developed by Maury in [122].

1.3 Linear analysis

1.3.1 Preliminaries

Under the assumptions considered in the previous section, the associated ODE systems are piecewise smooth. In particular, they are smooth in a neighbourhood of $\bar{\mathbf{X}}$, which allows us to carry out the linear stability analysis in order to study the local convergence of the solution towards a steady state.

We consider here the particular physical system where N rigid spheres in $\mathbb{R}^b$ attract each other through a global potential which is given by a quadratic function of the relative distance,

$$W(\mathbf{X}) = \frac{1}{2N} \sum_{i,j \in \{1,\dots,N\},\ i<j} |X_i - X_j|^2. \qquad (1.25)$$

Definition 14. *A steady state x^* of the ODE system $\dot{x} = f(x),\ t \geq 0$, is called*

- ***stable*** *(in the sense of Lyapunov) if for all $\epsilon > 0$, there exists a $\delta > 0$ such that $\|\bar{x}(0) - x^*\| < \delta$ implies $\|\bar{x}(t) - x^*\| < \epsilon$, for all $t > 0$ and for all solution $\bar{x}$;*
- ***asymptotically stable*** *if it is stable and $\lim_{t\to\infty} \|\bar{x}(t) - x^*\| = 0$;*
- ***unstable*** *if it is not stable.*

Note that this definition assumes that the initial configuration is chosen close enough to the steady state. Alternative notions of stability could have been used [151, 61], there is no particular reason for this choice. The next theorem allows us to obtain conclusions about the original nonlinear system from the corresponding linearized system.

Theorem 15. *Consider the ODE system $\dot{x} = f(x)$ and a steady state x^*, where f is smooth at x^*. If x^* is an asymptotically stable (unstable) solution of the linearized system about x^*, i.e., $\dot{\tilde{x}} = f'(x^*)(\tilde{x} - x^*)$, then it is an asymptotically stable (unstable) solution of the original system.*

Proof. See [34], thm. 2.42, p. 158. ∎

In order to ensure convergence of the ODE system towards a steady state, we only need to ensure that the eigenvalues of $f'(x^*)$ all have negative real part. If at least one eigenvalue has positive real part, then x^* is unstable, and if all eigenvalues are pure imaginary, then x^* is a *center equilibrium*, i.e. if a solution starts near it then it will be periodic around it. In the latter case, we cannot conclude anything about the nonlinear system. The analysis presented next is made for the case of two spheres in $\mathbb{R}$.

1.3.2 The Arrow-Hurwicz algorithm (AHA)

1.3.2.1 AHA-NS

Let $\phi(X) = d - |X|$ and consider the potential (1.25). The ODE system associated to the DAHA-NS in the case of two spheres in $\mathbb{R}$ where one sphere is fixed at the origin can be written as

$$\begin{cases} \dot{X} &= -\alpha \left(1 - \frac{\lambda}{|X|}\right) X & (1.26) \\ \dot{\lambda} &= \begin{cases} 0, \text{ if } \lambda = 0 \text{ and } d < |X| \\ \beta(d - |X|), \text{ otherwise.} \end{cases} & (1.27) \end{cases}$$

Lemma 16. *The steady states of the system* (1.26)-(1.27)*, $(X^*, \lambda^*) = (d, d)$ and $(X^*, \lambda^*) = (-d, d)$, are both asymptotically stable, for any α and β positive.*

Proof. Since the dynamics around each steady state is identical, we only need to carry out the analysis of the first steady state. Suppose $X > 0$ and consider the change of variables $Y = X - d$ and $\mu = \lambda - d$. The system on the new variables is given in matrix form by

$$\begin{bmatrix} \dot{Y} \\ \dot{\mu} \end{bmatrix} = A \begin{bmatrix} Y \\ \mu \end{bmatrix}, \; A = \begin{bmatrix} -\alpha & \alpha \\ -\beta & 0 \end{bmatrix}.$$

We want the eigenvalues of matrix A to be real and negative in order to have a fast convergence to the steady state. The roots of the characteristic polynomial $\mathcal{P}(\lambda) =$

$\lambda^2 + \alpha\lambda + \beta\alpha$, have both negative real part, therefore the steady state is asymptotically stable. ∎

Any solution to the ODE system (1.26)-(1.27) converges to a steady state for all $\alpha, \beta > 0$ and the fastest convergence is achieved when $\alpha = 4\beta$. In contrast to the one dimensional case, in higher spatial dimensions the constraints are no longer piecewise linear. Consequently, we cannot directly extrapolate the conclusions drawn in this section. In particular, in dimension $b = 2$, the numerical simulations show oscillations around the steady state for $N > 3$ without never converging to it. The non-convergence is in this case due to the non-existence of a saddle-point of the Lagrangian (see section 2.4.1).

1.3.2.2 AHA-S

Let $\phi(X) = d^2 - |X|^2$ and consider the potential (1.25). The ODE system associated to the AHA-S in the case of two spheres in $\mathbb{R}$ where one sphere is fixed at the origin can be written as

$$\begin{cases} \dot{X} & = -\alpha\,(1 - 2\lambda)\,X & (1.28) \\ \dot{\lambda} & = \begin{cases} 0, \text{ if } \lambda = 0 \text{ and } d < |X| \\ \beta(d^2 - X^2), \text{ otherwise.} \end{cases} & (1.29) \end{cases}$$

Lemma 17. *The steady states of the system corresponding to the linearization of* (1.28)-(1.29)*,* $(X^*, \lambda^*) = (d, 1/2)$ *and* $(X^*, \lambda^*) = (-d, 1/2)$*, are both center equilibria, for any* α *and* β *positive.*

Proof. As before, we will only carry out the analysis of the first steady state.

Suppose $X > 0$ and consider the change of variables $Y = X - d$ and $\mu = \lambda - 1/2$. The linearized system on the new variables is given in matrix form by

$$\begin{bmatrix} \dot{Y} \\ \dot{\mu} \end{bmatrix} = A \begin{bmatrix} Y \\ \mu \end{bmatrix}, \; A = \begin{bmatrix} 0 & 2d\alpha \\ -2d\beta & 0 \end{bmatrix}.$$

The roots of the characteristic polynomial $\mathcal{P}(\lambda) = \lambda^2 + 4d^2\alpha\beta$ are both purely imaginary, therefore the steady state of the linearized system is a center equilibrium. ∎

The linear analyses does not allow us to conclude anything about the asymptotic behaviour of the nonlinear system (see thm. 15). Nevertheless, the phase portrait plotted in Figure 1.2 reveals that a solution to the nonlinear system should converge towards a periodic orbit around the steady state. As we will see in the next section, the damping term applied to the Arrow-Hurwicz system (1.9)-(1.10) ensures asymptotic stability of the steady state, under certain conditions on the parameters.

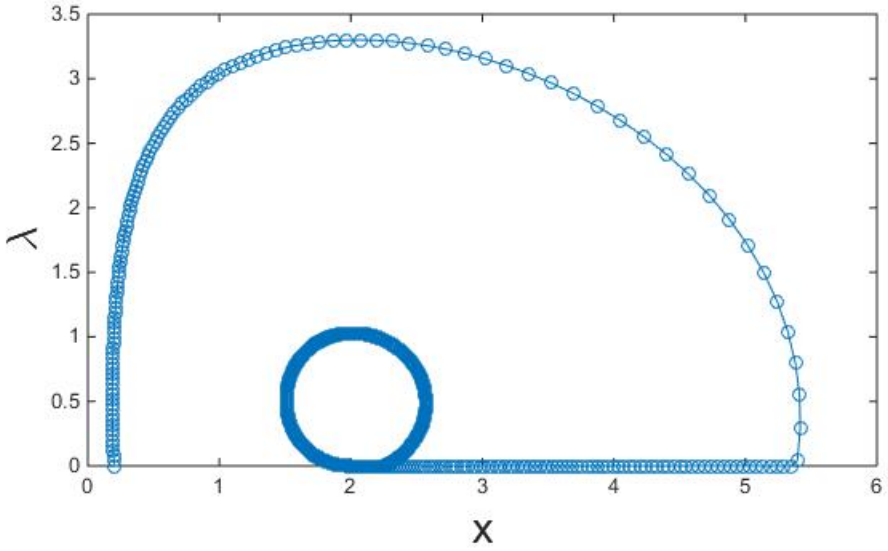


Figure 1.2: Phase portrait of the system (1.28)-(1.29) with $(\alpha, \beta, d) = (0.01, 0.01, 2)$ and initial condition $X^0 = 0.2$. The dynamics do not converge to the equilibrium $(2, \frac{1}{2})$.

1.3.3 The damped Arrow-Hurwicz algorithm (DAHA)

1.3.3.1 DAHA-NS

Let $\phi(X) = d - |X|$ and consider the potential (1.25). The ODE system associated to the DAHA-NS in the case of two spheres in $\mathbb{R}$ where one sphere is fixed at the origin can be written as

$$\begin{cases} \ddot{X} & = -\alpha^2 \left(1 - \frac{\lambda}{|X|}\right) X + \alpha\beta\lambda(d - |X|)\frac{X}{|X|} - c\dot{X} \qquad (1.30) \\ \dot{\lambda} & = \begin{cases} 0, \text{ if } \lambda = 0 \text{ and } d < |X|, \\ \beta(d - |X|), \text{ otherwise.} \end{cases} \qquad (1.31) \end{cases}$$

Lemma 18. *Let $\alpha, \beta, c > 0$. If $(\alpha + \beta d)c - \beta\alpha > 0$, then the steady states of the system (1.30)-(1.31), $(X^*, \dot{X}^*, \lambda^*) = (d, 0, d)$ and $(X^*, \dot{X}^*, \lambda^*) = (-d, 0, d)$, are both asymptotically stable.*

Proof. Suppose $X > 0$ and consider the change of variables $Y = X - d$, $Z = \dot{Y}$ and $\mu = \lambda - d$. The linearized system on the new variables is given in matrix form by

$$\begin{bmatrix} \dot{Y} \\ \dot{Z} \\ \dot{\mu} \end{bmatrix} = A \begin{bmatrix} Y \\ Z \\ \mu \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 & 0 \\ -\alpha^2 - \alpha\beta d & -c & \alpha^2 \\ -\beta & 0 & 0 \end{bmatrix}.$$

The eigenvalues of matrix A are the roots of the characteristic polynomial in λ, which is given by $\mathcal{P}(\lambda) = \lambda^3 + c\lambda^2 + (\alpha^2 + \alpha\beta d)\lambda + \beta\alpha^2$. Consider in general a cubic polynomial of the form $\mathcal{P}(\lambda) = \lambda^3 + c_2\lambda^2 + c_1\lambda + c_0$, with $c_0, c_1, c_2 \in \mathbb{R}^+$. Let z_1, z_2 and z_3 be the

(complex) roots to this polynomial. We want to ensure that all roots have negative real part. Since all coefficients are positive, if the roots are real then they must be negative. Suppose now that two roots are complex conjugate, for example, $z_1 = a + ib,\ z_2 = a - ib$, $a, b \in \mathbb{R}$ and $z_3 \in \mathbb{R}^-$. In order to find a condition on the coefficients which ensures that a is non-positive, we start by identifying the coefficients of the equation with its roots:

$$z_1 + z_2 + z_3 = -c_2,\ z_1z_2 + z_1z_3 + z_2z_3 = c_1,\ z_1z_2z_3 = -c_0$$

Rewriting in terms of a, b and z_3 we get

$$2a + z_3 = -c_2,\ a^2 + b^2 + 2az_3 = c_1,\ (a^2 + b^2)z_3 = -c_0 \tag{1.32}$$

From (1.32) we deduce that a satisfies the cubic polynomial

$$8a^3 + 8c_1a^2 + 2(c_1 + c_2^2)a + c_1c_2 - c_0 = 0.$$

Consequently, if $c_1c_2 - c_0 > 0$, then a is necessarily negative.

Back to our case, we have $c_2 = c,\ c_1 = \alpha^2 + \alpha\beta d$ and $c_0 = \beta\alpha^2$ and a sufficient condition for the steady state to be asymptotically stable is $(\alpha^2 + \alpha\beta d)c - \beta\alpha^2 > 0$, i.e., $(\alpha + \beta d)c - \beta\alpha > 0$. Note that since the steady state is asymptotically stable as a solution to the linearized system, then it is also asymptotically stable (see thm. 15). ∎

1.3.3.2 DAHA-S

Let $\phi(X) = d^2 - |X|^2$ and consider the potential (1.25). For the case of two spheres in $\mathbb{R}$ where one sphere is fixed at the origin, the ODE system associated to the DAHA-S can be written as

$$\begin{cases} \ddot{X} & = -\alpha^2\left(1 - 2\lambda\right)X + 2\alpha\beta\lambda(d^2 - |X|^2)X - c\dot{X} & (1.33) \\ \dot{\lambda} & = \begin{cases} 0,\ \text{if } \lambda = 0 \text{ and } d < |X| \\ \beta(d^2 - |X|^2),\ \text{otherwise.} \end{cases} & (1.34) \end{cases}$$

Lemma 19. *Let $\alpha, \beta, c > 0$. If $c - 2\alpha > 0$, then the steady states of the system (1.33)-(1.34), $(X^*, \dot{X}^*, \lambda^*) = (d, 0, 1/2)$ and $(X^*, \dot{X}^*, \lambda^*) = (-d, 0, 1/2)$, are both asymptotically stable.*

Proof. As before, suppose $X > 0$ and consider the change of variables $Y = X - d$, $Z = \dot{Y}$ and $\mu = \lambda - 1/2$. The linearized system on the new variables is given in matrix form by

$$\begin{bmatrix} \dot{Y} \\ \dot{Z} \\ \dot{\mu} \end{bmatrix} = A \begin{bmatrix} Y \\ Z \\ \mu \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 & 0 \\ -2\alpha\beta d^2 & -c & 2d\alpha^2 \\ -2d\beta & 0 & 0 \end{bmatrix}.$$

The eigenvalues of matrix A are the roots of the characteristic polynomial in λ:

$$\mathcal{P}(\lambda) = \lambda^3 + c\lambda^2 + 2\alpha\beta d^2\lambda + 4d^2\beta\alpha^2$$

Using the same reasoning as before we have $c_2 = c,\ c_1 = 2\alpha\beta d^2$ and $c_0 = 4d^2\beta\alpha^2$. A sufficient condition for the steady state to be asymptotically stable is $2c\alpha\beta d^2 - 4d^2\beta\alpha^2 > 0$, i.e., $c - 2\alpha > 0$. ∎

Remark 1.3.1. *We see that as long as the damping coefficient, c, is large enough, the sufficient conditions for stability of both the DAHA-NS and DAHA-S are fulfilled. Furthermore, the parameter space corresponding to the stability of DAHA-NS is larger than the one of the DAHA-S.*

1.3.4 Previous approaches

1.3.4.1 NAP-NS

Let $\phi(X) = d - |X|$ and consider its linearization around $X^p \neq 0$, i.e., $\phi^p(X) = d - \frac{X^p}{|X^p|}X$, and the potential (1.25). The ODE system associated to the inner-loop of the NAP-NS in the case of two spheres in $\mathbb{R}$ where one sphere is fixed at the origin can be written as

$$\begin{cases} \dot{X} &= -\alpha\left(X - \frac{X^p}{|X^p|}\lambda\right) & (1.35) \\ \dot{\lambda} &= \begin{cases} 0, \text{ if } \lambda = 0 \text{ and } d - \frac{X^p}{|X^p|}X < 0 \\ \beta(d - \frac{X^p}{|X^p|}X), \text{ otherwise,} \end{cases} & (1.36) \end{cases}$$

with the initial condition $(X, \lambda)(0) = (X^p, \lambda^p)$.

Lemma 20. *If $X^0 \neq 0$, then the steady state of the system (1.35)-(1.36), $(X^*, \lambda^*) = d\left(\frac{|X^p|}{X^p}, 1\right)$, is asymptotically stable for any α and β positive and the outer-loop converges in one iteration.*

Proof. The stability analysis shows that the steady state of (1.35)-(1.36), namely, $(X^*, \lambda^*) = d\left(\frac{|X^p|}{X^p}, 1\right)$, is asymptotically stable. Furthermore, if $X_0 \neq 0$, the outer-loop is defined recursively by $X^{p+1} = d\frac{|X^p|}{X^p}$ and it converges in one iteration. In fact, $X^1 = d\frac{|X^0|}{X^0}$ and $X_p = X_1$ for all $p > 1$. ∎

The conclusions of the analysis in the one dimensional case cannot be directly extrapolated to higher dimensional cases, where the constraint functions are no more piecewise linear. In section 1.4, we resort to numerical simulations to get some insight about the behaviour of the system in two spatial dimensions.

1.3.4.2 NAP-S

Let $\phi(X) = d^2 - |X|^2$ and consider its linearization around X^p, i.e., $\phi^p(X) = d^2 + |X^p|^2 - 2X^p \cdot X$ and the potential (1.25). The ODE system associated to the inner-loop of the NAP-S in the case of two spheres in $\mathbb{R}$ where one sphere is fixed at the origin can be written as

$$\begin{cases} \dot{X} & = -\alpha\,(X - 2X^p\lambda) \qquad (1.37) \\ \dot{\lambda} & = \begin{cases} 0, \text{ if } \lambda = 0 \text{ and } d^2 + (X^p)^2 - 2X^pX < 0 \\ \beta(d^2 + (X^p)^2 - 2X^pX), \text{ otherwise,} \end{cases} \qquad (1.38) \end{cases}$$

with the initial condition $(X, \lambda)(0) = (X^p, \lambda^p)$.

Lemma 21. *If $X^0 \neq 0$, then the steady state of the system (1.37)-(1.38), $(X^*, \lambda^*) = \frac{d^2+(X^p)^2}{2(X^p)^2}(X^p, \frac{1}{2})$, is asymptotically stable for any α and β positive and the outer-loop generates the sequence $\{X^p\}_{p\in\mathbb{N}}$ defined iteratively by $X^{p+1} = \frac{d^2+(X^p)^2}{2X^p}$, which is convergent.*

Proof. The stability analysis shows that the steady state of (1.37)-(1.38), namely, $(X^*, \lambda^*) = \frac{d^2+(X^p)^2}{2(X^p)^2}(X^p, \frac{1}{2})$, is asymptotically stable. Consequently, the outer-loop generates the sequence defined recursively by $X^{p+1} = \frac{d^2+(X^p)^2}{2X^p}$, which is well-defined for $X^0 \neq 0$. If this sequence is convergent to, say, L, then L must satisfy $L = \frac{d^2+L^2}{2L}$ i.e., $L = \pm d$. Now, if $X^p > d$ then

$$\frac{X^{p+1}}{X^p} = \frac{1}{2}\left(\frac{d^2}{(X^p)^2} + 1\right) < 1,$$

therefore $X^{p+1} < X^p$, i.e., the sequence decreases. Furthermore, if we write X^p in the form $X^p = d + \epsilon$, $\epsilon > 0$, we then have that

$$\begin{aligned} X^{p+1} - d &= \frac{1}{2}\left(\frac{d^2}{d+\epsilon} + d + \epsilon\right) - d \\ &= \frac{1}{2(d+\epsilon)}(d^2 + (d+\epsilon)^2 - 2d(d+\epsilon)) \\ &= \frac{\epsilon^2}{2(d+\epsilon)} > 0 \end{aligned}$$

i.e., $X^{p+1} > d$. On the other hand, if $0 < X^p < d$ then the sequence increases. Consequently, we finally conclude that if $X^0 > 0$ then the sequence $\{X^p\}_{p\in\mathbb{N}}$ converges towards d. Using the same reasoning we conclude that if $X^0 < 0$ then $\{X^p\}_{p\in\mathbb{N}}$ converges towards $-d$. We see that for the case of the smooth form of the constraint functions, the sequence generated by the outer-loop converges. ∎

1.3.4.3 NAV-NS

Let $\phi(X) = d - |X|$ and consider the linearization of ϕ around $X^p \neq 0$ evaluated at $X^p + \tau V$, i.e., $\phi^p(X^p + \tau V) = d - |X^p| - \tau \frac{X^p}{|X^p|} V$ and the potential (1.25). In the particular case of two spheres in $\mathbb{R}$ where one sphere is fixed at the origin, the ODE system corresponding to the inner-loop is given by

$$\begin{cases} \dot{V} & = -\alpha \left(V + X^p - \tau \ \lambda \frac{X^p}{|X^p|} \right) \qquad (1.39) \\ \dot{\lambda} & = \begin{cases} 0, \text{ if } \lambda = 0 \text{ and } d - |X^p| - \tau \frac{X^p}{|X^p|} V < 0 \\ \beta \left(d - |X^p| - \tau \frac{X^p}{|X^p|} V \right), \text{ otherwise,} \end{cases} \qquad (1.40) \end{cases}$$

with the initial condition $(V, \lambda)(0) = (V^p, \lambda^p)$.

Lemma 22. *If $X^0 \neq 0$, then the the steady state of the system (1.39)-(1.36) is asymptotically stable for any α and β positive and the outer-loop converges in one iteration.*

Proof. The stability analysis shows that the steady state,

$$V^* = \frac{d - |X^p|}{\tau} \frac{|X^p|}{X^p}, \ \lambda^* = \frac{1}{\tau^2}(d - |X^p|(1 - \tau)),$$

is asymptotically stable. Consequently, the outer-loop generates the sequence defined recursively by $X^{p+1} = d\frac{|X^p|}{X^p}$. If $X_0 \neq 0$ the sequence is well-defined and $X^1 = d\frac{|X^0|}{X^0}$ and $X_p = X_1$ for all $p > 1$, hence the outer-loop converges in one iteration. ∎

The NAV with the smooth form of the constraint functions did not show numerically good convergence results, for we do not explore it in this work. We note that in [122], the author has also only considered the non-smooth form of the constraint functions.

1.4 Numerical results

In this section we investigate and compare the numerical results obtained from the damped Arrow-Hurwicz algorithms (DAHA-NS, DAHA-S) and the nested algorithms (NAP-NS, NAP-S and NAV-NS) for the potential defined in (1.25). Due to the difficulty in finding the optimal parameters (α, β) for each method and for each N, we have restricted this study to the cases $N = 7$ and $N = 100$ in two spatial dimensions, $b = 2$. We address the convergence time, the robustness of the convergence time with respect to the initial configurations. Additionally, we compare the accuracy of the methods for the case $N = 7$ only. Indeed, in the case $N = 7$, the stable steady state of the dynamical systems associated to the algorithms is unique (apart from translations, rotations and reflections) and is represented in Figure 1.1b. This guarantees that all algorithms converge to the same minimum, whatever initial configuration. In particular, this allows us to assess the

accuracy of the algorithms by comparing the computed minimum with the exact one. We finally show some examples of configurations obtained with the DAHA-S for the case $N = 2000$ in two and three dimensions.

In order to adjust the spatial dimensions, the numerical parameters must satisfy $\alpha, \beta, c \sim \mathcal{O}(1)$ for the methods with the non-smooth form of the constraint functions and $\alpha, c \sim \mathcal{O}(1)$ and $\beta \sim \mathcal{O}(1/d^2)$ for the methods with the smooth form of the constraint functions. In the following we have considered $d = 1$.

In order to be able to compare the nested algorithms with the DAHA regarding convergence time, we only consider the evolution of $\mathbf{X}$ and $\mathbf{V}$ and we do not consider the evolution of $\boldsymbol{\lambda}$. We denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^{bN}$. For a given small and positive ϵ, the stopping criterion for the minimization algorithms associated to the NAP, is given by the following condition on the relative error

$$\frac{\|\mathbf{X}^{n+1} - \mathbf{X}^n\|}{\|\mathbf{X}^n\|} < \epsilon_{inner}. \tag{1.41}$$

For the case of the minimization problem formulated in terms of the velocities, the stopping criterion is similar but instead of $\mathbf{X}$ we write $\mathbf{V}$ and instead of normalizing by $\mathbf{V}^n$, we normalize by $\mathbf{X}^n$, yielding

$$\frac{\|\mathbf{V}^{n+1} - \mathbf{V}^n\|}{\|\mathbf{X}^n\|} < \frac{\epsilon_{inner}}{\tau}. \tag{1.42}$$

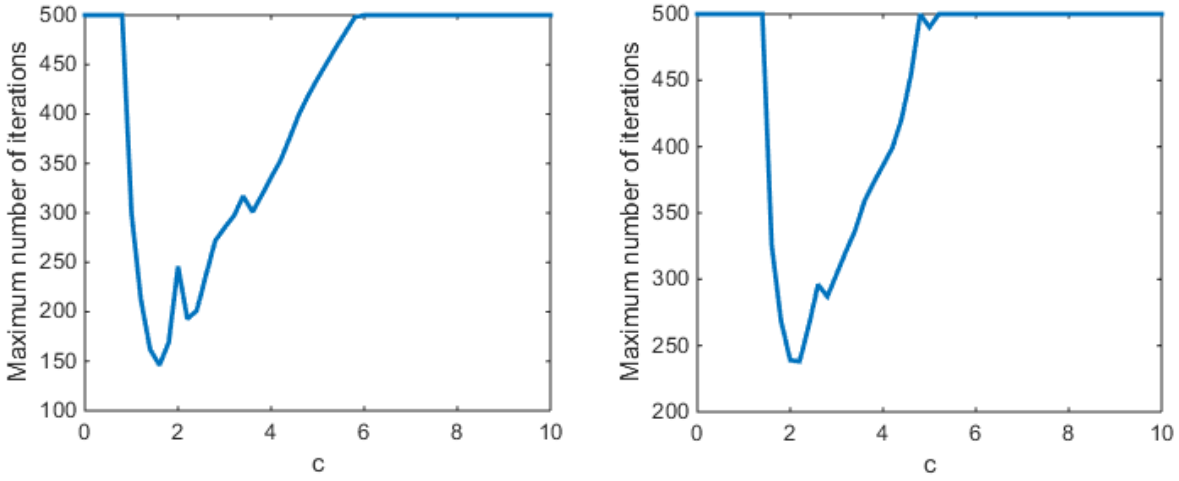
By using the Euler step $\mathbf{X}^{n+1} = \mathbf{X}^p + \tau\ \mathbf{V}^{n+1}$ we show that the two conditions (1.41) and (1.42) are equivalent. As we will see, in order to get a fast convergence with the nested algorithms, one does not need to wait for the convergence of the inner-loop. We introduce a new parameter, I_{inner}, which stands for the maximum number of iterations of the inner-loop allowed per outer-loop iteration. Finally, the stopping criterion for both the outer-loop of the NAP and the NAV, as well as, for the DAHA reads

$$\frac{\|\mathbf{X}^{p+1} - \mathbf{X}^p\|}{\|\mathbf{X}^p\|} < \epsilon. \tag{1.43}$$

The assessment and comparison of the methods will be made through the comparison of statistical indicators obtained from averaging certain quantities over a set of different initial configurations. The indicators are introduced bellow.

Definition 23. *Consider a set of p initial configurations for which an algorithm converges, i.e., the stopping criterion is satisfied in a finite number of iterations. Let T_ℓ be the number of iterations needed for the algorithm to converge when starting with the ℓth initial configuration. Let A_{ij} be the overlapping area of spheres i and j at convergence and $A_{total} = N\pi(d/2)^2$.*

We define the following statistical indicators **mean convergence time**, **variance of**



(a) DAHA-NS **(b)** DAHA-S

Figure 1.3: Maximum number of iterations needed for the DAHA to converge over a set of 20 randomly generated initial configurations as a function of c for $N = 7$ and $\epsilon = 10^{-6}$. The numerical parameters used are: 1.3a $(\alpha, \beta) = (0.3, 3)$ and 1.3b $(\alpha, \beta) = (0.3, 1.4)$.

the convergence time *and the* **mean proportion of overlapping area per sphere** *as*

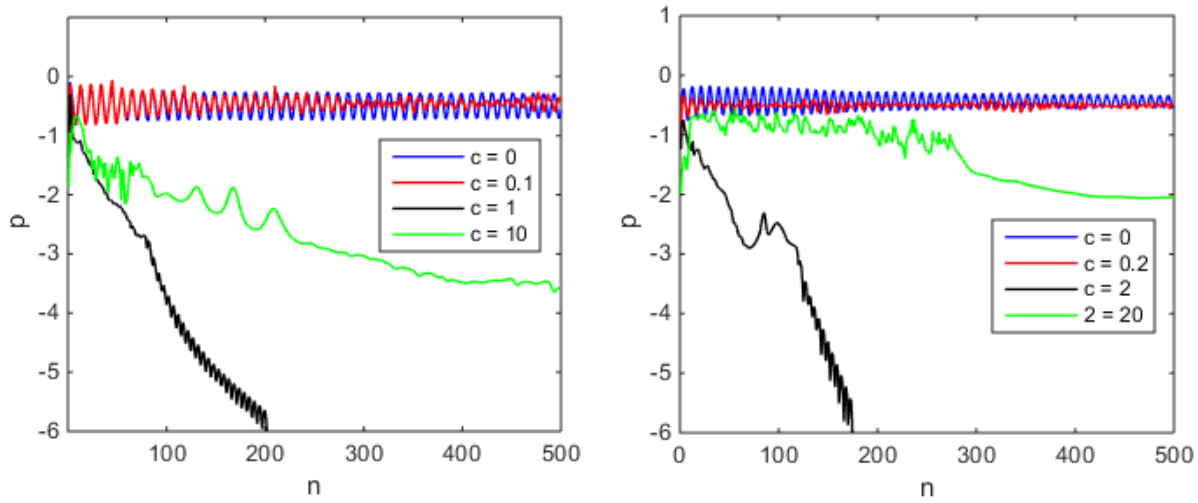
$$T = \frac{1}{p}\sum_{\ell=1}^{p} T_\ell, \quad \sigma^2 = \frac{1}{p-1}\sum_{\ell=1}^{p}(T_\ell - T)^2 \quad \textit{and} \quad A = \frac{1}{pNA_{total}} \sum_{i,j\in\{1,\dots,N\},\ i<j} A_{ij},$$

respectively.

The indicator T measures the efficiency of an algorithm with respect to the convergence time, A measures the accuracy on the treatment of the non-overlapping constraints and σ^2 measures the robustness of the convergence time with respect to the initial configurations. For simplicity we assume that the time interval between iterations is constant and invariant among the different algorithms. As a consequence of this simplification, we will use the number of iterations as the time unit of T.

1.4.1 Case $N = 7$

We present a detailed numerical study for the case of $N = 7$ spheres in dimension $b = 2$. The 20 different initial configurations considered in this section were generated from a standard Gaussian distribution. We choose the tolerances $\epsilon = 10^{-6}$ and $\epsilon_{inner} = 10^{-9}$ and the maximum number of iterations of the inner-loop $I_{inner} = 10$. In order to study the relation between the damping parameter c and the convergence time of the DAHA with smooth and non-smooth constraints, we plot in Figure 1.3 the maximum number of iterations over 20 different randomly generated initial configurations as a function of $c \in (0, 10]$. We observe that the lower convergence time is attained when $c \approx 2$, for both



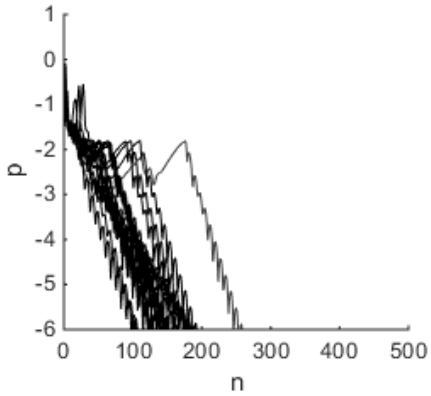
(a) DAHA-NS **(b)** DAHA-S

Figure 1.4: Relative error on log scale averaged over a set of 20 randomly generated initial configurations, p, as a function of iteration number, n, for different values of c and for $N = 7$ and $\epsilon = 10^{-6}$. The numerical parameters used are: 1.4a $(\alpha, \beta) = (0.3, 3)$ and 1.4b $(\alpha, \beta) = (0.35, 1.4)$.

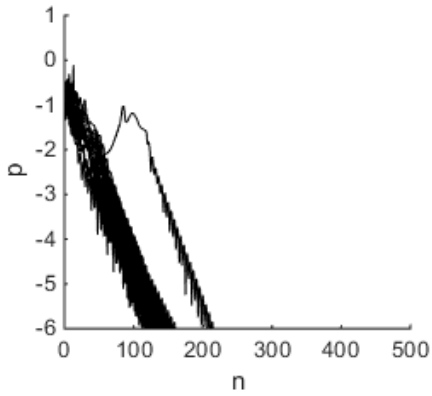
the DAHA with the smooth and with the non-smooth constraints. In Figure 1.4 we plot the relative error as a function of iteration number, n, for different values of c. If $c = 0$ we observe that the relative error oscillates and never drops bellow 10^{-1}. As we increase c the oscillations tend to diminish. In the following we have used $c = 2$. Note that this choice for c eliminates the dependence on $\mathbf{X}^{n-1}$ in (1.14)-(1.15), in this case, the DAHA can be seen as a discretization of the following first-order ODE system:

$$\begin{cases} \dot{X}_i &= -\frac{1}{2}\alpha^2[\nabla_{X_i} W(\mathbf{X}) + \sum\limits_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X})] \\ & \quad -\frac{1}{2}\alpha\beta \sum\limits_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X})\lambda_{k\ell}\nabla_{X_i}\phi_{k\ell}(\mathbf{X}),\ i = 1,\dots,N \\ \dot{\lambda}_{k\ell} &= \begin{cases} 0, & \text{if } \lambda_{k\ell} = 0 \text{ and } \phi_{k\ell}(\mathbf{X}) < 0 \\ \beta\phi_{k\ell}(\mathbf{X}), & \text{otherwise} \end{cases}, \\ & \qquad\qquad k,\ell = 1,\dots,N,\ k < \ell. \end{cases}$$

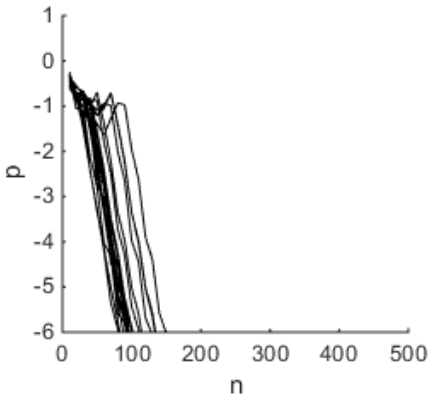
We now consider the five methods, namely, the DAHA-NS, DAHA-S, NAP-NS, NAP-S and NAV-NS. The numerical simulations suggest that α has influence in the attraction and β in the repulsion between spheres, which can also be observed by looking directly at the equations. If β is too large we observe oscillations, if α or both parameters are too large we observe numerical instability and if α is small we observe a very slow dynamics. The optimal set of parameters should then be chosen near the parameters that lead to numerical instability, which may be a problem, since an algorithm may be very efficient for some set of initial configurations, however it may be numerically unstable for another. Moreover, we have observed that the final positions obtained from each method by varying the numerical parameters are naturally not always the same. Only the relative positions



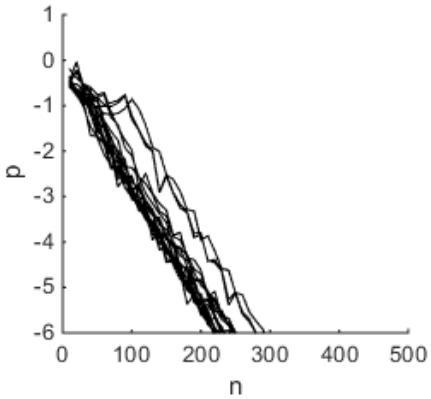
(a) DAHA-NS



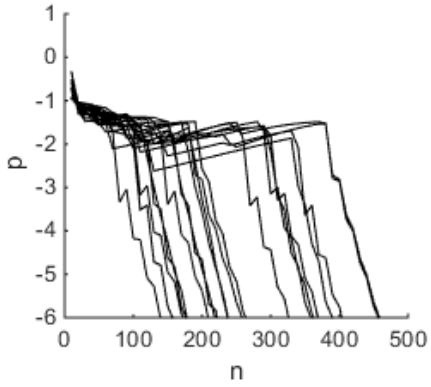
(b) DAHA-S



(c) NAP-NS



(d) NAP-S



(e) NAV-NS

Figure 1.5: Relative error on log scale, p, as a function of iteration number, n, for 20 randomly generated initial configurations computed by each algorithm for $N = 7$ and $\epsilon = 10^{-6}$. The numerical parameters used are: 1.5a $(\alpha, \beta, c) = (0.3, 3, 2)$, 1.5b $(\alpha, \beta, c) = (0.35, 1.4, 2)$, 1.5c $(\alpha, \beta, I_{inner}, \epsilon_{inner}) = (0.6, 0.46, 10, 10^{-9})$, 1.5d $(\alpha, \beta, I_{inner}, \epsilon_{inner}) = (0.25, 0.28, 10, 10^{-9})$ and 1.5e $(\alpha, \beta, I_{inner}, \epsilon_{inner}, \tau) = (0.48, 126, 10, 10^{-9}, 0.1)$.

$(\epsilon, \epsilon_{inner})$	$(10^{-4}, 10^{-7})$	$(10^{-6}, 10^{-9})$	$(10^{-8}, 10^{-11})$
Mean convergence time (T)			
DAHA-NS	103	145	198
DAHA-S	85	133	182
NAP-NS	**81**	**107**	**131**
NAP-S	165	246	325
NAV-NS	248	283	535
Order of accuracy $(A, W-3)$			
DAHA-NS	$(10^{-6}, -10^{-6})$	$(10^{-9}, -10^{-9})$	$(10^{-12}, -10^{-13})$
DAHA-S	$(10^{-7}, -10^{-6})$	$(10^{-10}, -10^{-10})$	$(10^{-12}, -10^{-13})$
NAP-NS	$(\mathbf{10^{-9}, 10^{-8}})$	$(\mathbf{10^{-12}, 10^{-12}})$	$(\mathbf{10^{-15}, 0})$
NAP-S	$(10^{-8}, -10^{-5})$	$(10^{-11}, 10^{-6})$	$(10^{-13}, -10^{-8})$
NAV-NS	$(10^{-8}, 10^{-1})$	$(10^{-10}, -10^{-1})$	$(10^{-14}, 10^{-9})$
Variance of the convergence time (σ^2)			
DAHA-NS	1.08×10^3	1.21×10^3	1.70×10^3
DAHA-S	4.19×10^2	4.49×10^2	5.65×10^2
NAP-NS	$\mathbf{3.00 \times 10^2}$	$\mathbf{4.13 \times 10^2}$	$\mathbf{2.93 \times 10^2}$
NAP-S	5.74×10^2	4.68×10^2	4.68×10^2
NAV-NS	1.03×10^4	1.04×10^4	1.80×10^5

Table 1.1: Results of the assessment of the final configurations averaged over a set of 20 initial configurations and obtained by each algorithm for $N = 7$ and for three different values of ϵ, namely, 10^{-4}, 10^{-6} and 10^{-8}. The parameters used are DAHA-NS $(\alpha, \beta, c) = (0.3, 3, 2)$, DAHA-S $(\alpha, \beta, c) = (0.35, 1.4, 2)$, NAP-NS $(\alpha, \beta, I_{inner}, \epsilon_{inner}) = (0.6, 0.46, 10, 10^{-9})$, NAP-S $(\alpha, \beta, I_{inner}, \epsilon_{inner}) = (0.25, 0.28, 10, 10^{-9})$ and NAV-NS $(\alpha, \beta, I_{inner}, \epsilon_{inner}, \tau) = (0.48, 126, 10, 10^{-9}, 0.1)$.

(apart from permutations) are invariant. As an exception, the NAV leads to nearly exactly the same configurations. In fact, we can see from equations (1.2)-(1.3) that the dynamics of the particles in the NAV is not directly affected by a change of the numerical parameters, since the parameters are only involved in the computation of the velocity. In contrast, in the NAP and DAHA a change in the parameters produces a different dynamics, which leads to different final configurations. This is comprehensible, since the parameters in those algorithms are directly related to the attraction and repulsion forces between the spheres. The previous observation could be statistically verified by comparing all the final configurations produced by each method and checking how different (how far away from each other) they are.

In the following, we have chosen the numerical parameters (α, β) that correspond to a fast convergence of each method for all the 20 sets of initial configurations. We plot in Figure 1.5 the relative error on log scale as a function of iteration number for 20 randomly generated initial configurations and for each method. We observe that the profile of the relative error follows the pattern: non-monotone behaviour, followed by an approximately linear decay at a certain speed, which seems to be invariant with respect to the initial configuration. The faster decay is observed in the NAP-NS (see Figure 1.5c). The efficiency of the NAV-NS (see Figure 1.5e) is apparently highly dependent on the initial configuration.

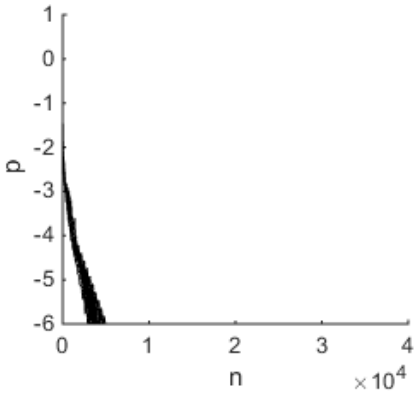
In order to quantify and compare the efficiency, as well as, the accuracy of the final configurations generated by each algorithm we consider three different tolerances, $\epsilon = 10^{-4}$, $\epsilon = 10^{-6}$ and $\epsilon = 10^{-8}$, and we compute the mean convergence time, T, the variance of the convergence time, σ^2, the mean proportion of overlapping area per sphere A and the difference between the theoretical optimum and the value of W at convergence. The results are presented in Table 1.1 and they are averaged over a set of 20 randomly generated initial configurations. We observe that the NAP-NS (in bold) performs better than any other method, while the NAV-NS is the least robust to initial configurations, the slowest to reach convergence and it only produces an accurate solution for $\epsilon = 10^{-8}$. In the case of the methods with smooth constraints, we also observe in Table 1.1 that the DAHA-S converges faster and produces more accurate solutions than the NAP-S. Indeed, for $\epsilon = 10^{-6}$ for example, we observe that the mean convergence time is $T = 133$ for the DAHA-S and $T = 246$ for the NAP-S and the difference between the theoretical minimum and the computed one is of order 10^{-10} for the DAHA-S and 10^{-6}) for the NAP-S.

1.4.2 Case $N = 100$

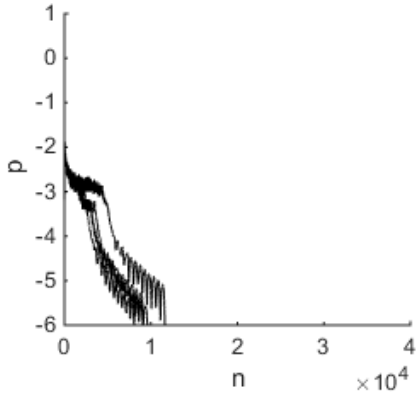
We present in the following a short study for the case of $N = 100$ spheres. The 5 different initial configurations considered in this section were generated from a standard Gaussian distribution. In this case where the initial configurations are very dense, one may observe two different types of behaviours depending on the choice of the numerical parameters: either the spheres disperse initially very rapidly before they start to concentrate again while trying to avoid overlapping with other spheres or they disperse slowly while trying to rearrange in a non-overlapping configuration. We keep the choice $\epsilon = 10^{-6}$ and $\epsilon_{inner} = 10^{-9}$, and we choose $I_{inner} = 10^3$. Similarly to the case $N = 7$, the best value for the damping parameter c should be of order $\mathcal{O}(1)$. We keep the choice $c = 2$.

We now consider the five methods, namely, the DAHA-NS, DAHA-S, NAP-NS, NAP-S and NAV-NS. In the following we have chosen the numerical parameters (α, β) that correspond to a fast convergence of each method for all the 5 sets of initial configurations. We plot in Figure 1.6 the relative error on log scale as a function of iteration number. In contrast to the case $N = 7$ the DAHA seems to converge faster than the NAP (see Figures 1.6a-1.6d). The efficiency of the nested algorithms (see Figures 1.6c-1.6e) seems to be highly dependent on the initial configuration.

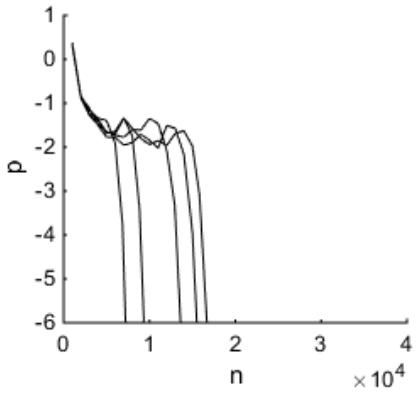
Note that the performance of the methods depends not only on the numerical parameters, but also on the initial configuration. In this study, we have only considered initial configurations that are very concentrated around one point. In the other case, i.e., if the spheres are initially far away from each other, then all the simulations must be redone and different conclusions may be drawn.



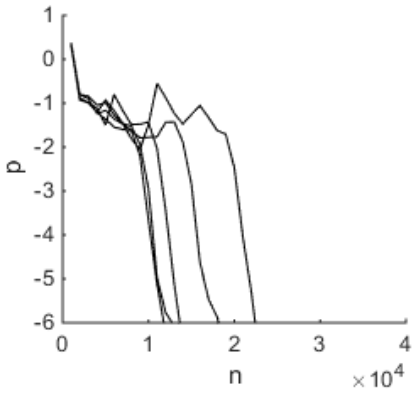
(a) DAHA-NS



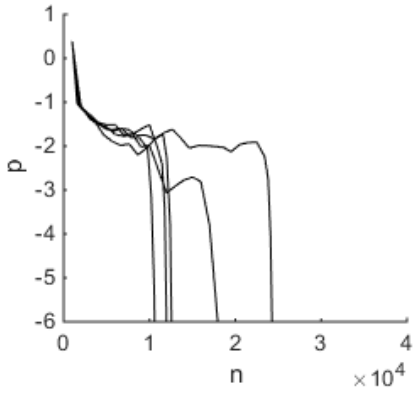
(b) DAHA-S



(c) NAP-NS



(d) NAP-S



(e) NAV-NS

Figure 1.6: Relative error on log scale, p, as a function of iteration number, n, for 5 randomly generated initial configurations computed by each algorithm for $N = 100$ and $\epsilon = 10^{-6}$. The numerical parameters used are: 1.6a $(\alpha, \beta, c) = (0.07, 0.5, 2)$, 1.6b $(\alpha, \beta, c) = (0.04, 0.15, 2)$, 1.6c $(\alpha, \beta, I_{inner}, \epsilon_{inner}) = (0.1, 0.16, 10^3, 10^{-9})$, 1.6d $(\alpha, \beta, I_{inner}, \epsilon_{inner}) = (0.015, 0.026, 10^3, 10^{-9})$ and 1.6e $(\alpha, \beta, I_{inner}, \epsilon_{inner}, \tau) = (0.31, 41, 10^3, 10^{-9}, 0.1)$.

1.4.3 Case $N = 2000$

Numerical simulations were successfully performed with the DAHA for N large, up to $N = 2000$. In order to obtain a faster convergence towards a local minimizer, a fourth parameter, γ, was introduced in the first equation of the original formulation of the DAHA (1.14)-(1.15), yielding,

$$\begin{cases} X_i^{n+1} = \frac{1}{1+c/2}\left(2X_i^n - (1-c/2)X_i^{n-1}\right) \\ \qquad -\frac{\alpha^2}{1+c/2}[\nabla_{X_i} W(\mathbf{X}^n) + \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \lambda_{k\ell}^n \nabla_{X_i}\phi_{k\ell}(\mathbf{X}^n)] \\ \qquad -\frac{\gamma^2}{1+c/2} \sum_{k,\ell\in\{1,\dots,N\},\ k<\ell} \phi_{k\ell}(\mathbf{X}^n)\lambda_{k\ell}^n \nabla_{X_i}\phi_{k\ell}(\mathbf{X}^n),\ i = 1,\dots,N \\ \lambda_{k\ell}^{n+1} = \max\{0, \lambda_{k\ell}^n + \beta\phi_{k\ell}(\mathbf{X}^{n+1})\},\ k,\ell = 1,\dots,N,\ k<\ell. \end{cases}$$

While using $\gamma \neq \sqrt{\alpha\beta}$ cannot be justified from the derivation presented in section 1.2.4, it seems to bring additional flexibility to the algorithm that can be used to improve speed and accuracy (see Chapter 2).

Note that the parameter γ is imperative for the dynamics to converge. If we set $\gamma = 0$ with $c = 2$, the DAHA is reduced to the AHA (up to a scaling factor) and the algorithm presents an oscillatory behaviour as observed in figure 1.2.

In Figure 1.7 we present an example of four configurations that were obtained at intermediate steps, namely, $n = 1, n = 101, n = 1001$ and $n = 10001$ for $N = 2000$ with the DAHA-S. The initial configuration was generated from a standard Gaussian distribution. For a tolerance of $\epsilon = 10^{-5}$ an optimal configuration was obtained in less than 13000 iterations.

We observe numerically that the stationary state is close to an optimal packing configuration, which corresponds to the hexagonal lattice. In order to quantify this observation, we measure the packing density, denoted by ϕ, which corresponds to the ratio of the area covered by the particles over a disc centered at the center of mass $\bar{X} = \frac{1}{N}\sum_{i=1,\dots,N} X_i$ for different radius R (see Figure 1.8). We observe that the packing density ($\phi \approx .87$) is close to the optimal packing density (i.e. $\frac{\pi}{2\sqrt{3}} \approx .909$) for a wide range of radius size R. For $R > 25$, the packing density quickly decays since there are not enough particles to cover the domain considered.

1.4.4 Case $N = 2000$ in $\mathbb{R}^3$

Finally, we would like to explore how the DAHA performs in $\mathbb{R}^3$. We use a similar setting as in $\mathbb{R}^2$, i.e., we choose the same parameters with an initial configuration generated from a standard Gaussian distribution. In figure 1.9, we plot the configuration after 10^4 iterations and remove a segment to visualize its interior. The color corresponds to the pressure exerted by nearby particles and estimate as $\mu_i = \sum_j \lambda_{ij}$. Notice that

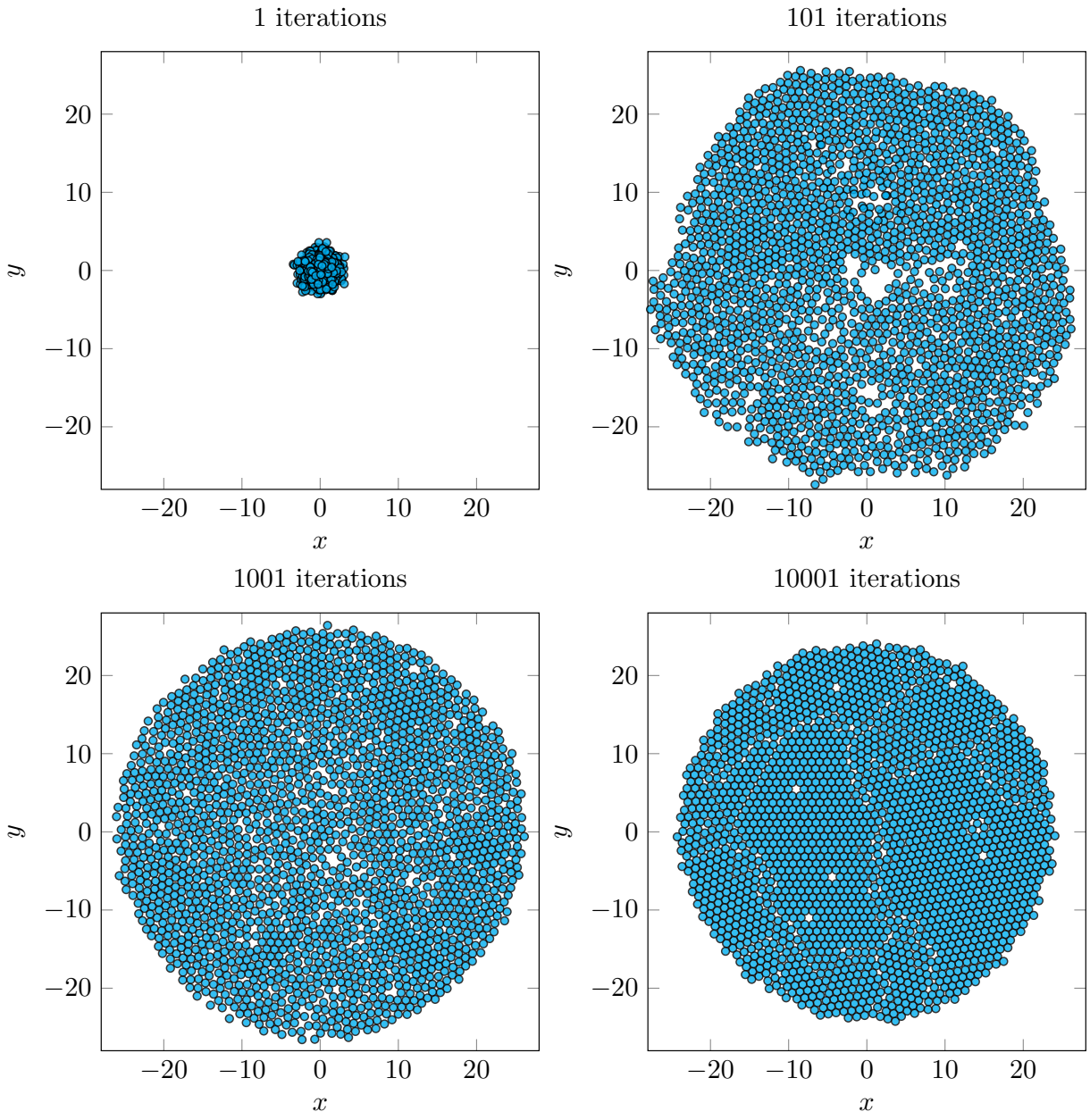
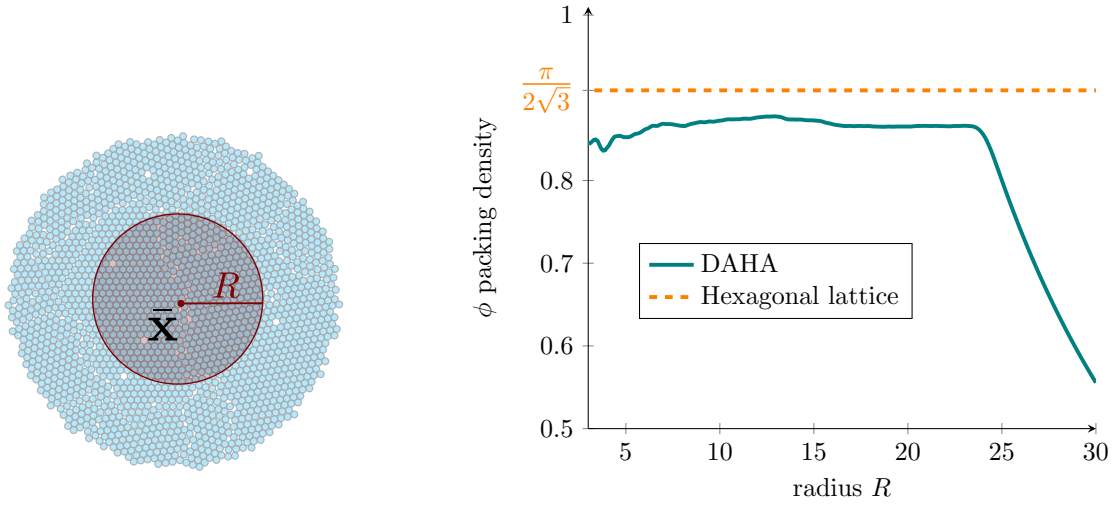


Figure 1.7: Sequence of configurations produced at intermediate steps, namely, $n = 1$, $n = 101$, $n = 1001$ and $n = 10001$ with the DAHA-S and for $N = 2000$. The numerical parameters used are: $(\alpha, \beta, \gamma, c) = (7.8 \times 10^{-3}, 2.8 \times 10^{3}, 1.6 \times 10^{-2}, 2)$.



(a) **(b)**

Figure 1.8: The packing density ϕ is estimated by taking the ratio of the area covered by the particles over a disc of size R centered at $\bar{X} = \frac{1}{N}\sum_{i=1,\dots,N} X_i$, with $N = 2000$ (1.8a). The packing density is close to the optimal configuration for a wide range of radius R (1.8b).

numerically the dynamics has not yet reached a stationary state but it would require a new investigation to analyze when equilibrium will be reached.

We estimate the density packing of the configuration using the same method as in $\mathbb{R}^2$ except that the domain considered is a ball (centered at the center of mass $\bar{X} = \frac{1}{N}\sum_{i=1,\dots,N} X_i$ with radius R) instead of a disc. In figure 1.10a, we observe that the packing density ϕ reaches a maximum around 0.65. The value $\phi \approx 0.64$ is actually the packing density for a random close packing [173]. However, the optimal configuration for sphere packing would give a packing density of $\phi = \frac{\pi}{3\sqrt{2}} \approx 0.741$. It is an open problem to determine if one could get closer to this optimal value by using a different set of parameters.

Another useful information is to study the number of neighbours each particle has. Numerically, two particles are neighbours if their relative distance is less than 1.1 (1 being the distance for two discs in contact). In figure 1.10b, we observe that particles have mainly between $8-10$ neighbours. In an optimal packing configuration, each particle would have 12 neighbours.

1.4.5 Summary of the results

We confront in table 1.2 the results obtained from the theoretical analysis for the case of two spheres ($N = 2$) in one dimension ($b = 1$) and the results of the numerical simulations for the case of $N > 2$ spheres in two dimensions ($b = 2$). If the system converges to a non-overlapping configuration within a reasonable number of iterations and for some set of parameters we write ✓, otherwise we write ✗.

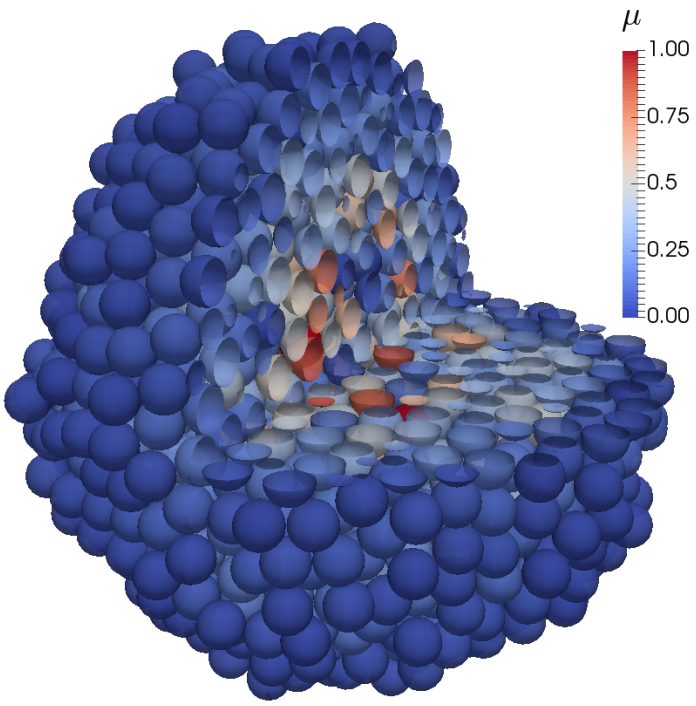
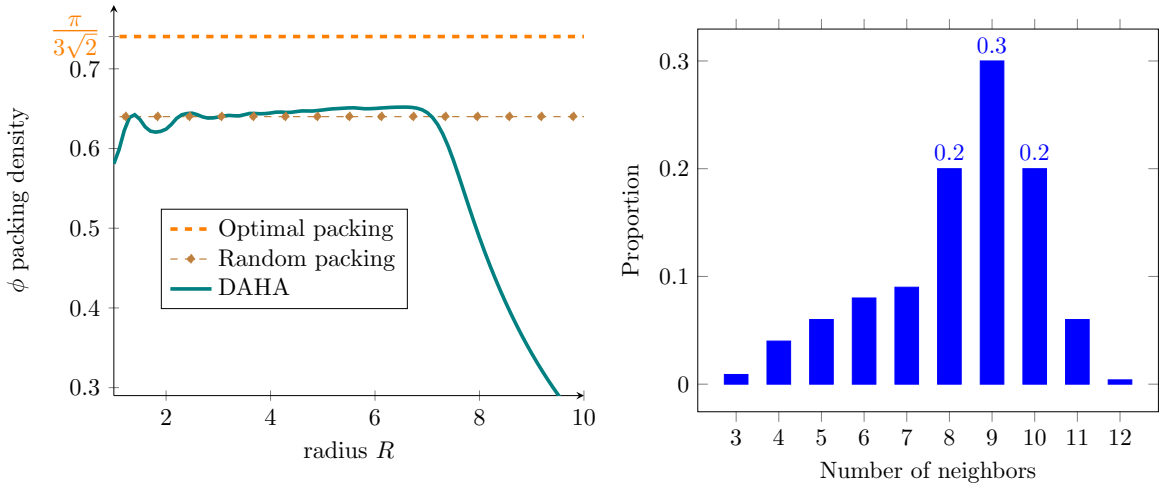


Figure 1.9: Configuration after $n = 10^4$ iterations of the DAHA-S algorithm in $\mathbb{R}^3$ with $N = 2000$ particles. The color corresponds to the pressure exerted by nearby particles denoted μ. The numerical parameters used are: $(\alpha, \beta, \gamma, c) = (2.8 \times 10^{-2}, 2.2 \times 10^{2}, 5.6 \times 10^{-2}, 2)$.



(a) **(b)**

Figure 1.10: Packing density for the three dimensional configuration of Figure 1.9 (1.10a). Histogram of the number of neighbours for all particles (1.10b).

	Smooth constraints		Non-smooth constraints	
	Analysis	Simulations	Analysis	Simulations
AHA	✗	✗	✓	✗
DAHA	✓	✓	✓	✓
NAP	✓	✓	✓	✓
NAV	✓	✗	✓	✓

Table 1.2: Summary of the results obtained from the analysis for $N = 2$ in one spatial dimension and numerical simulations for $N \geq 2$ in two spatial dimensions.

1.5 Conclusions and future work

We have deduced a promising algorithm for solving a non-convex minimization problem, which was derived from a multi-step variant of the Arrow-Hurwicz algorithm: the damped Arrow-Hurwicz algorithm. This algorithm can be seen as a generalization of the AHA, when an additional parameter, γ, is considered. In the particular case of packing problems, the DAHA has revealed to perform better for a large number of spheres when compared to other classical algorithms. Further studies should be done in order to explore both the advantages and limitations of this method. In Chapter 2 we develop the stability analysis of the corresponding ODE system for the case of a generic number of spheres N in $b = 1$ dimension. However, the analysis in the general case of higher dimensions is still missing, as well as, the analysis of the numerical stability. In the present Chapter, the DAHA was assessed in the case of a global potential and highly dense initial configurations, which we believe to correspond to the worst scenario possible. Nevertheless, the results obtained here do not necessarily apply to other types of potentials or initial configurations, and hence similar studies should be conducted for those cases. The estimated packing density of a three dimensional configuration obtained with the DAHA shows that this algorithm is capable of generating random close packings. Whether a higher density could be achieved by choosing a different set of parameters remains an open problem. In the next work we should consider more general particle systems with different sized spheres or ellipsoids. The applications of hard-particle systems are vast, for we believe these type of algorithms are going to be very useful in the study of many biological, physical and social systems. As an example of application to the modelling of a cell tissue see Chapter 4.

2

Analytical study of a non-convex minimization problem

We investigate a non-convex minimization problem with non-overlapping constraints which naturally arises in the study of many-particle systems with congestion (see Section 0.3.1.2). Specifically, we consider algorithms to numerically solve this problem which are based on the Lagrangian formulation of the minimization problem. We first prove that the Lagrangian does not have any saddle-point which implies the non-convergence of the classical Arrow-Hurwicz algorithm. We then study an Augmented Lagrangian method and a first-order version of the damped Arrow-Hurwicz algorithm proposed in Chapter 1. We prove that these two methods converge to the only solution (apart from permutations and translations) to the minimization problem in dimension $d = 1$.

This work is based on a collaboration with Sebastien Motsch[*]*. The proof of the non-existence of a saddle-point presented in Section 2.4.1 has been contributed by Sebastien and it has been included in this thesis for completeness.*

[*]School of Mathematical & Statistical Sciences, Arizona State University, Tempe, USA

2.1 Introduction

Minimization problems under non-overlapping constraints occur naturally in the modelling of systems of interacting particles (see Section 0.3.1.2). The constraints are associated to the shape of the particles, which is typically described by inequalities involving non-convex functions, consequently, leading to non-convex optimization. Unfortunately there are no algorithms that work for all non-convex problems, therefore they have to be treated case by case. The aim of this Chapter is to study several algorithms to numerically solve these type of problems. In particular we formalize some of the numerical results obtained in Chapter 1 in terms of the properties of solutions to the minimization problem and convergence of the algorithms.

Given a configuration of N spheres of radius R in $\mathbb{R}^d$, $d \in \mathbb{N}$, we are interested in finding an *optimal packing configuration* that is close to the given one. To formulate this problem, we denote $\mathbf{x}_i$ the center of each sphere and $X = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ the vector of the configuration and we consider the potential W:

$$W(X) = \frac{1}{2N} \sum_{i<j} |\mathbf{x}_i - \mathbf{x}_j|^2, \tag{2.1}$$

where $|\cdot|$ denotes Euclidean norm, and the non-overlapping constraints :

$$\phi_{ij} \leq 0 \quad \text{for all } i \neq j \quad \text{with } \phi_{ij} = 4R^2 - |\mathbf{x}_i - \mathbf{x}_j|^2. \tag{2.2}$$

We combine these $N(N-1)/2$ constraints into a single vector: $F(X) = (\phi_{ij})_{i<j}$. The expression for the potential W (2.1) has been chosen based on what we expect to be the worst scenario possible in terms of convergence of the algorithms considered in this Chapter. Indeed, the convergence corresponds to a balance between attraction and repulsion forces. In the case of the potential (2.1) this balance is particularly difficult to achieve for large N due to the large magnitude of attraction forces, which require a fast actualization of the repulsion forces. We claim that if the algorithms perform well with this potential, then they will also perform well for other convex potentials. This claim should be further investigated.

Definition 24. *A configuration $\overline{X}$ is an optimal packing configuration if it satisfies locally:*

$$\overline{X} \in \underset{F(X)\leq 0}{\operatorname{argmin}} W(X) \tag{2.3}$$

where $F(X) \leq 0$ indicates that each component of F should be non-positive (i.e. $\phi_{ij} \leq 0$ for all $i \neq j$). In other words, $\overline{X}$ minimizes the potential W under the non-overlapping constraints.

We are interested in solving the optimization problem (2.3). We illustrate the problem

with two solutions in dimension 2 and 3 (see Figure 2.1). The main challenge of the problem (2.3) is that the constraints do **not** form a convex domain. In other words, the following set is not convex:

$$\mathcal{C} = \{X : \ F(X) \leq 0\} = \{X : \ |\mathbf{x}_j - \mathbf{x}_i| \geq 2R \text{ for all } i \neq j\}. \tag{2.4}$$

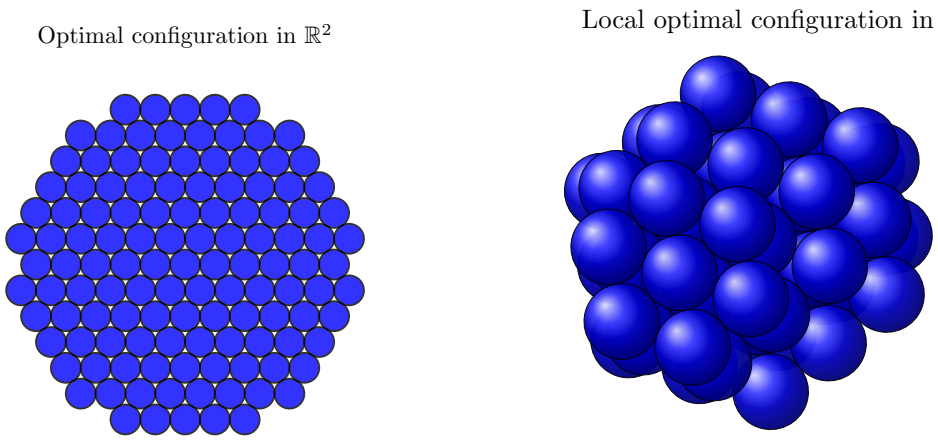


Figure 2.1: Two examples of solutions to the packing problem (2.3). A global optimum in dimension 2 is presented on the left and corresponds to the hexagonal lattice. A local optimum in dimension 3 is presented on the right.

Since the problem (2.3) is invariant under translation, we can restrict the set of solutions for vector X with zero mean. Under this restriction, the potential W is coercive:

$$W(X) \overset{|X|\to+\infty}{\longrightarrow} +\infty.$$

Since the domain $\mathcal{C}$ is closed, we deduce that there always exists a solution.

Proposition 25. *There exist solutions $\overline{X}$ to the problem* (2.3).

Of course, the solution is far from being unique. From invariance by rotation and translation, one can deduce other solutions. But for dimensions bigger than 1 there are in general even more solutions. In appendix A we present some solutions in dimension $d = 2$ and for $N = 2,\ 3,\ 4$.

Suppose that $\overline{X} = (\mathbf{x}_i)_{i=1,\dots,N}$ is a solution of (2.3) and that the Abadie Constraint qualification (ACQ) [2] holds at $\overline{X}$ (see Section 1.2.2). Then, by Theorem 4 there exist Lagrange multipliers $\overline{\lambda} = (\lambda_{ij})_{i<j}$ satisfying the three conditions: the feasibility conditions,

$$\lambda_{ij} \geq 0, \ F(\overline{X}) \leq 0 \tag{2.5}$$

the complementary slackness conditions,

$$\lambda_{ij}\phi_{ij}(\overline{X}) = 0. \tag{2.6}$$

and the Euler-Lagrange equation,

$$\nabla_X W(\overline{X}) + \sum_{i<j} \lambda_{ij} \nabla_X \phi_{ij}(\overline{X}) = 0. \tag{2.7}$$

Substituting (2.1) and (2.2) in (2.7), we get for each i:

$$\frac{1}{N} \sum_j (\mathbf{x}_i - \mathbf{x}_j) - 2 \sum_j \lambda_{ij} (\mathbf{x}_i - \mathbf{x}_j) = 0. \tag{2.8}$$

A Lagrange multiplier λ_{ij} is non-zero only if the corresponding constraint is *activated*:

$$\lambda_{ij} > 0 \quad \Rightarrow \quad |\mathbf{x}_i - \mathbf{x}_j| = 2R,$$

Conditions (2.5)-(2.7) are called Karush-Kuhn-Tucker conditions (KKT-conditions) and they constitute the necessary conditions for optimality [134]. These conditions are on the basis of the optimization algorithms treated in this Chapter.

The outline of the Chapter is as follows. We start by presenting in Section 2.2 the dynamical systems associated to the algorithms considered in this Chapter, namely, the Arrow-Hurwicz algorithm (AHA), the augmented Lagrangian algorithm (ALA) and the first-order damped Arrow-Hurwicz algorithm (DAHA), and we summarize the convergence results towards a solution of the minimization problem. In Section 2.3 we find the only analytical solution, apart from permutations and translations of the minimization problem (2.3) in dimension $d = 1$. We then compute the order of magnitude of the Lagrange multipliers associated to what we call a spherical configuration for any dimension. These results are used in the investigation of the algorithms that are developed in Section 2.4. In particular, in Section 2.4.1 we show that in general that the Lagrangian associated to the minimization problem does not have any saddle-point, which implies the non-convergence of the classical Arrow-Hurwicz algorithm. Then in Section 2.4.2 we analyse two algorithms in dimension $d = 1$ which are based on modifications of the Arrow-Hurwicz algorithm and show that they converge to a solution of the minimization problem for any positive value of the parameters, provided the initial configuration is chosen close enough to the solution. Finally, in Section 2.5 we draw some conclusions and future directions.

2.2 Algorithms and convergence results

In this Section we present the dynamical systems associated to each algorithm treated in this Chapter and the convergence results towards a solution of the minimization problem (2.3).

2.2.1 Lagrangian method and Arrow-Hurwicz algorithm (AHA)

A classical method to solve the minimization problem (2.3) is based on the so-called Lagrangian $\mathcal{L} : \mathbb{R}^{dN} \times (\mathbb{R}_0^+)^{N(N-1)} \to \mathbb{R}$,

$$\begin{aligned} \mathcal{L}(X, \lambda) &= W(X) + \langle \lambda, F(X) \rangle \\ &= \frac{1}{2N} \sum_{i<j} |\mathbf{x}_j - \mathbf{x}_i|^2 + \sum_{i<j} \lambda_{ij} \phi_{ij}, \end{aligned} \tag{2.9}$$

with $\phi_{ij} = 4R^2 - |\mathbf{x}_j - \mathbf{x}_i|^2$. From equations (2.5)-(2.7) we see that a solution to the minimization problem (2.3) is a *critical point* to $\mathcal{L}$. We now introduce the concept of *saddle point*.

Definition 26. *A couple (X_*, λ_*) is a saddle point of the Lagrangian $\mathcal{L}$ (2.9) if it satisfies:*

$$\mathcal{L}(X_*, \lambda) \leq \mathcal{L}(X_*, \lambda_*) \leq \mathcal{L}(X, \lambda_*) \tag{2.10}$$

for any X and $\lambda \geq 0$.

One can show if (X_*, λ_*) is a saddle point of $\mathcal{L}$ then it also satisfies the minimization problem (2.3). In other words, being a saddle point for $\mathcal{L}$ is a *stronger* condition than solving the minimization problem (2.3) (consequence of the weak-duality). In some special cases, such as linear constraints or convex problems with Slater's condition (see Section 0.3.3.1), the two problems are equivalent (strong-duality [27]). But it will not be the case in our setting of sphere packing, there is a *duality gap*.

To reach the saddle point of $\mathcal{L}$, we introduce the Arrow-Hurwicz system [17, 184] (AHS):

$$\begin{cases} \dot{X} &= -\alpha \nabla_X \mathcal{L} \\ \dot{\lambda} &= \beta \nabla_\lambda \mathcal{L}. \end{cases} \tag{2.11}$$

where α, β are two positive numbers. Using expression (2.9) of the Lagrangian $\mathcal{L}(X, \lambda)$, this leads to:

$$\begin{cases} \dot{\mathbf{x}}_i &= \dfrac{\alpha}{N} \displaystyle\sum_{j=1}^{N} (\mathbf{x}_j - \mathbf{x}_i) \;+\; 2\alpha \sum_{j=1}^{N} \lambda_{ij} (\mathbf{x}_i - \mathbf{x}_j) \\ \dot{\lambda}_{ij} &= \begin{cases} 0, & \text{if } \lambda_{ij} = 0 \text{ and } \phi_{ij} < 0 \\ \beta \phi_{ij}, & \text{otherwise.} \end{cases} \end{cases} \tag{2.12}$$

with $\phi_{ij} = 4R^2 - |\mathbf{x}_j - \mathbf{x}_i|^2$. The Arrow-Hurwicz algorithm (AHA) corresponds to a time-discretization of this system. In all simulations presented in this Chapter we use a semi-implicit Euler discretization, specifically, we use an explicit Euler discretization in the first equation for X and implicit Euler discretization in the second equation for λ.

In a convex setting, one can find sufficient conditions to show the convergence of the Arrow-Hurwicz method [184]. However, it will not be possible in our setting. Indeed,

as we will prove in the next Section, there is **no** saddle point for the Lagrangian $\mathcal{L}$ (see Theorem 34).

Remark 2.2.1. *Since we are interested in finding the equilibrium of the system* (2.11)*, we can change the time unit and reduce the number of parameters. Indeed, using as new time variable* $\tilde{t} = \sqrt{\alpha\beta}\ t$ *and considering the change* $\tilde{\lambda} = \sqrt{\alpha}/\sqrt{\beta}\lambda$ *and* $\varepsilon = \sqrt{\alpha}/\sqrt{\beta}$*, we obtain after dropping the tildes:*

$$\text{(AHS)} \quad \begin{cases} \dot{\mathbf{x}}_i &= \dfrac{\varepsilon}{N}\displaystyle\sum_{j=1}^{N}(\mathbf{x}_j - \mathbf{x}_i) \ + \ 2\sum_{j=1}^{N}\lambda_{ij}(\mathbf{x}_i - \mathbf{x}_j) \\ \dot{\lambda}_{ij} &= \begin{cases} 0, & \text{if } \lambda_{ij} = 0 \text{ and } \phi_{ij} < 0 \\ \phi_{ij}, & \text{otherwise.} \end{cases} \end{cases} \tag{2.13}$$

Notice that we could in principle get rid of the parameter ε *using a change in unit length. However, length is already prescribed by the radius* R *of the spheres. Thus, modifying unit length will modify the radius* R*. In the following we will consider this version of the Arrow-Hurwicz system* (2.13)*, which is based on the modified Lagrangian* $\mathcal{L}_\varepsilon$*:*

$$\begin{aligned} \mathcal{L}_\varepsilon(X,\lambda) &= \varepsilon W(X) + \langle \lambda, F(X)\rangle \\ &= \frac{\varepsilon}{2N}\sum_{i<j}|\mathbf{x}_j - \mathbf{x}_i|^2 + \sum_{i<j}\lambda_{ij}\phi_{ij}, \end{aligned} \tag{2.14}$$

In this case, the Euler-Lagrange equation (2.8) *becomes*

$$\frac{\varepsilon}{N}\sum_j(\mathbf{x}_i - \mathbf{x}_j) - 2\sum_j \lambda_{ij}(\mathbf{x}_i - \mathbf{x}_j) = 0. \tag{2.15}$$

As we will observe and prove later, the Arrow-Hurwicz method does not converge for our non-convex setting. Therefore we study two alternatives which are also based on the Lagrangian associated to the minimization problem. These two alternatives are presented in the two following Sections.

2.2.2 Augmented Lagrangian Algorithm (ALA)

We propose to add an extra term to the Arrow-Hurwicz system (2.13) to obtain what we call the augmented Lagrangian system (ALS) as follows:

$$\text{(ALS)} \quad \begin{cases} \dot{\mathbf{x}}_i &= \frac{\varepsilon}{N}\sum_{j=1}^{N}(\mathbf{x}_j - \mathbf{x}_i) \ + \ 2\sum_j \left(\lambda_{ij} + \gamma\phi_{ij}\right)^+ (\mathbf{x}_i - \mathbf{x}_j) \\ \dot{\lambda}_{ij} &= \begin{cases} 0, & \text{if } \lambda_{ij} = 0 \text{ and } \phi_{ij} < 0 \\ \phi_{ij}, & \text{otherwise.} \end{cases} \end{cases} \tag{2.16}$$

where γ and ε are positive parameters and $x^+ = \max(x, 0)$ represents the positive part of a real number x. The augmented Lagrangian algorithm (ALA) is obtained by a semi-

implicit Euler discretization of the ALS, specifically, we use an explicit Euler discretization in the first equation for X and an implicit Euler discretization in the second equation for λ. Note that for $\gamma = 0$ we retrieve the original Arrow-Hurwicz system.

We observe that any steady state $(\bar{X}, \bar{\lambda})$ of the system (2.16) satisfies the complementary slackness condition (2.6). Hence, $(\bar{X}, \bar{\lambda})$ is also a steady state to the Arrow-Hurwicz system (2.13) and a critical point to the original Lagrangian $\mathcal{L}$. Thus, the modification in (2.16) does not change the equilibrium states.

In dimension $d = 1$, system (2.16) has only one steady state, apart from permutations and translations, and in Section 2.3.1 we characterize this state. In Section 2.4.2 we prove that this steady state is Lyapunov asymptotically stable, i.e., for any initial configuration chosen close enough to the steady state, the system will converge to it for any positive value of the parameters.

Remark 2.2.2. *Note that the algorithm* (2.16) *is close to an augmented Lagrangian method [134], the algorithm adds up a memory term λ_{ij} to the strength of repulsion γ. Thus, if a constraint is not satisfied i.e. $\phi_{ij} > 0$, the coefficient λ_{ij} will increase and make the repulsion between i and j. Therefore, there is no need to take $\gamma \to +\infty$.*
However, classical augmented Lagrangian would be slightly different. Indeed, if we consider the following Augmented Lagrangian:

$$\widetilde{\mathcal{L}_\varepsilon}(X, \lambda) = \frac{\varepsilon}{2N} \sum_{i<j} |\mathbf{x}_j - \mathbf{x}_i|^2 \; + \; \sum_{i<j} \lambda_{ij} \phi_{ij} \; + \; \gamma \sum_{i<j} (\phi_{ij}^+)^2. \tag{2.17}$$

Applying Arrow-Hurwicz method leads to (assuming $\alpha = 1$):

$$\begin{cases} \dot{\mathbf{x}}_i &= \dfrac{\varepsilon}{N} \displaystyle\sum_{j=1}^{N} (\mathbf{x}_j - \mathbf{x}_i) \; + \; 2 \sum_j \left(\lambda_{ij} + \gamma \phi_{ij}^+\right) (\mathbf{x}_i - \mathbf{x}_j) \\ \dot{\lambda}_{ij} &= \begin{cases} 0, & \text{if } \lambda_{ij} = 0 \text{ and } \phi_{ij} < 0 \\ \phi_{ij}, & \text{otherwise.} \end{cases} \end{cases} \tag{2.18}$$

Thus, we obtain as repulsion coefficient: $\left(\lambda_{ij} + \gamma \phi_{ij}^+\right)$ rather than $(\lambda_{ij} + \gamma \phi_{ij})^+$ in (2.16). *Numerically, the algorithm* (2.16) *outperforms the algorithm* (2.18).

2.2.3 First-order damped Arrow-Hurwicz algorithm (DAHA)

We also investigate the first-order damped Arrow-Hurwicz system that has been obtained from the original second-order damped Arrow-Hurwicz system proposed in [45] by considering a certain parameter regime. This version of the method has revealed to be the most appropriate to solve problem (2.3) (see Section 1.4.1). The first-order damped Arrow-

Hurwicz system (DAHS) is given by:

$$\text{(DAHS)} \quad \begin{cases} \dot{\mathbf{x}}_i &= \dfrac{\varepsilon}{N}\displaystyle\sum_{j=1}^{N}(\mathbf{x}_j - \mathbf{x}_i) \;+\; 2\sum_j (\lambda_{ij} + \gamma\lambda_{ij}\phi_{ij})(\mathbf{x}_i - \mathbf{x}_j) \\ \dot{\lambda}_{ij} &= \begin{cases} 0, & \text{if } \lambda_{ij} = 0 \text{ and } \phi_{ij} < 0 \\ \phi_{ij}, & \text{otherwise.} \end{cases} \end{cases} \tag{2.19}$$

The first-order damped Arrow-Hurwicz algorithm (DAHA) is obtained by a semi-implicit Euler discretization of the DAHS, specifically, we use an explicit Euler discretization in the first equation for X and an implicit Euler discretization in the second equation for λ.

This system is very similar to the augmented Lagrangian system. The only difference is the γ term, which has the additional factor λ_{ij} in the case of the first-order DAHS. By the same reason as in the previous Section, the equilibrium states correspond to the critical-points of the Lagrangian.

Similarly to the ALS, we prove in Section 2.4.2 that, in dimension $d = 1$, this system converges to the only solution (apart from permutations and translations) of the minimization problem, provided the initial configuration is close enough to the steady state. Numerical experiments performed in Chapter 1 indicate that in higher dimensions $d = 2$ and $d = 3$ solutions to the minimization problem also correspond to asymptotically stable steady states of this system, however, a proof is still missing.

The two systems ALS and DAHS differ on the rate of convergence towards the steady state as a function of parameter space, as we may conclude from the analysis developed in Section 2.4.2. The best parameters in terms of speed of convergence towards the steady state, lead to a faster convergence of the DAHS when compared to the ALS. On the other hand, such region within the parameter space is wider in the case of the ALS than in the case of the DAHS.

2.3 Properties of analytical solutions

2.3.1 Solution in dimension $d = 1$

In this Section we present an explicit formula for the only solution apart from permutations and translations to the minimization problem (2.3) in dimension $d = 1$. This result is going to be used later in Section 2.4.2 in the study of convergence of the algorithms AL and DAHA.

Proposition 27. *In dimension $d = 1$, the configuration $\overline{X} = (x_1, ..., x_N)$ defined by*

$$\sum_j x_j = 0 \text{ and } x_{i+1} = x_i + 2R, \;\; i = 1, ..., N-1$$

is the only solution (apart from permutations) to (2.3). *Moreover, the associated Lagrange multipliers* $\overline{\lambda} = (\lambda_{ij})_{i,j}$ *are unique and given by* $\lambda_{i,j} = 0,\ j \neq i-1, i+1$ *and*

$$\lambda_{N-k,N-k+1} = \begin{cases} \varepsilon \frac{2Mk-(k-1)k}{4}, & \text{for } N = 2M+1 \\ \varepsilon \frac{2Mk-k^2}{4}, & \text{for } N = 2M \end{cases}, \ k = 1, ..., N-1$$

Proof. The configuration $\overline{X}$ is an admissible point to (2.3) as

$$\phi_{i,i+1}(\overline{X}) = 0,\ i = 1, ..., N-1, \quad \phi_{i,j}(\overline{X}) = -(i-j+1)^2 4R^2 < 0,\ i \neq j-1, j+1 \tag{2.20}$$

and it corresponds to segments arranged next to each other (see Figure 2.2). Consider

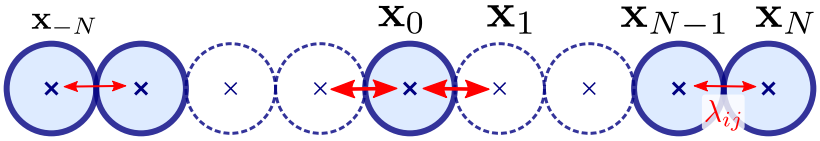


Figure 2.2: Explicit solution to the minimization problem (2.3) in dimension 1.

an arbitrary perturbation $\Upsilon = (\upsilon_1, ..., \upsilon_N) \in \mathbb{R}^N \backslash \{0\}$ such that $\overline{X} + \Upsilon$ is an admissible point centered at 0, i.e., $\sum_i (x_i + \upsilon_i) = 0$, and the order of the segments is preserved (no permutations), ie, $x_i + \upsilon_i < x_{i+1} + \upsilon_{i+1}$. Consequently, $\overline{X} + \Upsilon$ satisfies the following conditions:

$$\phi_{i,i+1}(\overline{X} + \Upsilon) = 4R^2 - (x_i + \upsilon_i - x_{i+1} - \upsilon_{i+1})^2 \leq 0,\ i = 1, ..., N-1, \tag{2.21}$$

$$\sum_i \upsilon_i = 0, \tag{2.22}$$

$$\upsilon_i - \upsilon_{i+1} < 4R,\ i = 1, ..., N-1. \tag{2.23}$$

From condition (2.21) we get $4R(\upsilon_i - \upsilon_{i+1}) \leq (\upsilon_i - \upsilon_{i+1})^2$ and together with (2.23) implies that $\upsilon_i - \upsilon_{i+1}$ has to be non-positive. Therefore $\upsilon_i \leq \upsilon_j$ for all $i < j$. Moreover, if $\upsilon_i = \upsilon_j$ for all i, j, then from (2.22) we would have $\Upsilon = 0$ which is impossible, consequently there exists at least one pair (i, j) for which the inequality is strict. This implies

$$W(\overline{X} + \Upsilon) - W(\overline{X}) = \sum_{i<j} (\upsilon_i - \upsilon_j)^2 - 4R \sum_{i<j} (\upsilon_i - \upsilon_j) > 0,$$

which proves that $\overline{X}$ is the only solution to (2.3) apart from permutations.

We now compute the Lagrange multipliers $\overline{\lambda}$ associated to $\overline{X}$. Suppose that $N = 2M+1$. Applying formula (2.15) for $i = N$, we obtain:

$$\varepsilon M \cdot 2R = 2\lambda_{M-1,M}(2R) \quad \Rightarrow \quad \lambda_{M-1,M} = \frac{\varepsilon M}{2}.$$

For $i = N-1$, we deduce:

$$\varepsilon(M-1)\cdot 2R = 2\big(\lambda_{M-2,M-1}\cdot 2R - \lambda_{M-1,M}\cdot 2R\big) \quad \Rightarrow \quad \lambda_{M-2,M-1} = \frac{\varepsilon(M-1)}{2} + \frac{\varepsilon M}{2}.$$

Thus, recursively, we obtain: $\lambda_{N-k,N-k+1} = \frac{\varepsilon(M-k)}{2} + \cdots + \frac{\varepsilon M}{2}$, ie,

$$\lambda_{N-k,N-k+1} = \varepsilon\frac{2Mk-(k-1)k}{4}.$$

A similar argument applies also with $N = 2M$ spheres. Finally, from (2.6) we get that $\lambda_{i,j} = 0$ for $j \neq i-1,\ i+1$, which concludes the proof. ∎

Remark 2.3.1. *For N odd in Prop. 27, the maximum of the second-order polynomial in k is reached at $k = M + 1/2$. The closest integers are $k = M$ and $k = M+1$, which corresponds to the middle of the configuration. At those points the value of λ is $\lambda_{M,M+1} = \lambda_{M+1,M+2} = \varepsilon\frac{M(M+1)}{4}$.*

2.3.2 Spherical solution

In dimension $d = 2$, one can find analytically the minimizer $\overline{X}$ and the corresponding Lagrange multipliers with $N \leq 4$ spheres (see appendix A). However for larger values of N or d the problem becomes much harder and there is in general no uniqueness of $\overline{X}$ nor explicit expression for the Lagrange multipliers. It is though possible to deduce the order of magnitude of the Lagrange multipliers $\overline{\lambda}$ with respect to N associated to a particular type of solution belonging to the set of spherical configurations that we define next. These results are used in the study of the existence of saddle-points of the Lagrangian presented in the Section 2.4.1.

Given a configuration $X = (\mathbf{x}_i)_{i=1,\dots,N}$ in $\mathbb{R}^d$, let G be the undirected graph associated to X where the vertices $V = \{1,\dots,N\}$ correspond to the indexes of the points in X and the edges $E = \{(i,j)|\phi_{ij}(X) = 0\}$ correspond to the pairs of neighbouring spheres. A path P between any two vertices $i, j \in V$ corresponds to a sequence of different vertices where each vertex is connected to the following one, ie, $P = (v_1,\dots,v_n) \in V^n$, with $(v_i, v_{i+1}) \in E$ and $v_i \neq v_j,\ i \neq j$. The set of all paths of G from v_i to v_f is denoted by $\mathcal{P}_{v_f}^{v_i}(G)$. We define the length of a path P as a function $L : \mathcal{P}_{v_i}^{v_f}(G) \to \mathbb{N},\ P = (v_1,\dots,v_n) \mapsto L(P) = n-1$. We define a distance f between any two vertices of a connected graph G as

$$f(v_i, v_f) = \min_{P\in\mathcal{P}_{v_i}^{v_f}(G)} L(P). \tag{2.24}$$

For technical purposes, suppose that there exists one sphere $\mathbf{x}_1$ located at the center of the configuration, ie., $\mathbf{x}_1 = \langle\mathbf{x}\rangle := 1/N\sum_i \mathbf{x}_i$. Now, consider a partition of X defined

by the length of the shortest path to the center of the configuration $\langle \mathbf{x} \rangle$, ie,

$$X = \bigcup_{i=1}^{M} \chi_i, \tag{2.25}$$

with

$$\chi_i = \{\mathbf{x}_k \in \overline{X} | f(k,1) = i-1\}, \quad i = 1, ..., M. \tag{2.26}$$

Note that each set corresponds to a level set of the distance $f(\cdot, 1)$ to the center.

Definition 28. *Let $X = \{\mathbf{x}_i\}_{i=1,...,N}$ be a configuration in $\mathbb{R}^d$ of spheres with radii of order 1 with respect to N. Consider the partition of X defined in (2.24)-(2.26). We say that X is a spherical configuration with radius $\mathcal{R} := \max_k |\mathbf{x}_k - \langle \mathbf{x} \rangle|$ if it satisfies the non-overlapping constraints (2.2) and the two conditions:*

$$\textit{(convexity)} \qquad \mathcal{O}(|\mathbf{x}_i - \mathbf{x}_j|) = f(i,j), \quad \textit{for all } i,j = 1, ..., N \tag{2.27}$$

$$\textit{(sphericity)} \qquad \mathcal{O}(|\mathbf{x}_k - \langle \mathbf{x} \rangle|) = i, \quad \textit{for all } \mathbf{x}_k \in \chi_i \textit{ and } i = 1, ..., M. \tag{2.28}$$

where $\mathcal{O}$ denotes order of magnitude.

Remark 2.3.2.

- *The spherical configuration is an admissible point and should have an approximately spherical shape with radius $\mathcal{R} := \max_k |\mathbf{x}_k - \langle \mathbf{x} \rangle| = \mathcal{O}(M)$ as shown in the examples of Figure 2.1.*
- *This definition is based on the idea of convexity and on the idea of defining a sphere as the union of spherical surfaces successively contained in each other, ie., the level curves of $f(1,k) = \min_{P \in \mathcal{P}_{x_k}^{\langle x \rangle}(G)} L(P)$ are approximately contained in spherical surfaces in $\mathbb{R}^d$ with center $\langle x \rangle$.*

Two examples of spherical configurations are presented in Figure 2.1. We start by computing the order of magnitude of the radius of a spherical configuration.

Lemma 29. *Let X be a spherical configuration with radius $\mathcal{R}$ of N spheres in $\mathbb{R}^d$. Then $\mathcal{O}(\mathcal{R}) = N^{1/d}$.*

Proof. The volume V of the d-dimensional sphere of radius $\mathcal{R}$ has order of magnitude of $\mathcal{R}^d$. Since V has the same order of magnitude as the sum of the volume of the spheres that constitute X, we have $\mathcal{O}(V) = N$. Consequently, $\mathcal{O}(\mathcal{R}^d) = N$, ie., $\mathcal{O}(\mathcal{R}) = N^{1/d}$. ∎

Similarly to the one dimensional case, we show that the Lagrange multipliers λ_{ij} are increasing as we move towards the center of the configuration.

Proposition 30. *Let X be a solution to the minimization problem (2.3) in dimension d that belongs to the set of spherical configurations. Moreover, let λ be the associated Lagrange multipliers. Then, the order of magnitude of λ_{ij} ranges from $\varepsilon N^{1/d}$ to $\varepsilon N^{2/d}$, for spheres i and j at the periphery and at the center of the configuration, respectively.*

Proof. Consider the partition of X in M sets as defined in (2.24)- (2.26). For any sphere k in χ_k, let $k_j,\ j = 1, ..., n_k, ..., \overline{n}_k$ be its neighbouring spheres, such that (see Figure 2.3)

$$\begin{cases} (\mathbf{x}_{k_j} - \mathbf{x}_k) \cdot (\langle \mathbf{x} \rangle - \mathbf{x}_k) \leq 0, & j = 1, ..., n_k - 1 \\ (\mathbf{x}_{k_j} - \mathbf{x}_k) \cdot (\langle \mathbf{x} \rangle - \mathbf{x}_k) \geq 0, & j = n_k, ..., \overline{n}_k \end{cases} \tag{2.29}$$

Denote

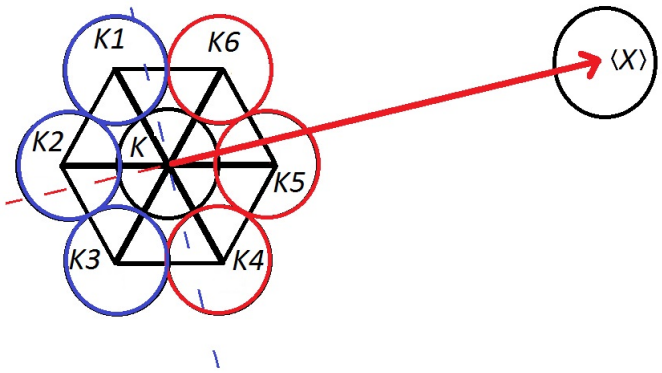


Figure 2.3: Sketch of the interaction between spheres k, its neighbours $k_1, ..., k_6$ and $\langle \mathbf{x} \rangle$ in dimension 2. The dot product between vector $\langle \mathbf{x} \rangle - \mathbf{x}_k$ and vectors $\mathbf{x}_{k_j} - \mathbf{x}_k$ is negative for $j = 1, 2, 3$ (circles in blue) and positive for $j = 4, 5, 6$ (circles in red).

$$\sigma_k^- := \sum_{j=1}^{n_k - 1} \lambda_{kk_j} \text{ and } \sigma_k^+ := \sum_{j=n_k}^{\overline{n}_k} \lambda_{kk_j}. \tag{2.30}$$

We want to prove by induction that

$$\sigma_k^+ = \mathcal{O}\left(\varepsilon(M + M - 1 + ... + k)\right),\ k = 1, ..., M, \tag{2.31}$$

where $k = M$ represents a sphere at the periphery and $k = 1$ at the center of the configuration.

We first introduce some notation and deduce some relations. Consider the angle θ_{kk_j} between the vectors $\mathbf{x}_{k_j} - \mathbf{x}_k$ and $\langle \mathbf{x} \rangle - \mathbf{x}_k$ and define the non-negative constants $\alpha_{kk_j} := |\cos(\theta_{kk_j})|$. Moreover, considering the projection onto $\mathbf{x}_k - \langle \mathbf{x} \rangle$ we have from equation (2.15) that:

$$\varepsilon|\mathbf{x}_k - \langle \mathbf{x} \rangle|^2 = 2 \cdot 2R|\mathbf{x}_k - \langle \mathbf{x} \rangle|(-\sum_{j=1}^{n_k - 1} \lambda_{kk_j}\alpha_{kk_j} + \sum_{j=n_k}^{\overline{n}_k} \lambda_{kk_j}\alpha_{kk_j}).$$

Equivalently,

$$\varepsilon|\mathbf{x}_k - \langle \mathbf{x} \rangle| = 2 \cdot 2R(-\sum_{j=1}^{n_k - 1} \lambda_{kk_j}\alpha_{kk_j} + \sum_{j=n_k}^{\overline{n}_k} \lambda_{kk_j}\alpha_{kk_j}).$$

Since the coefficients α_{kk_j} have order 1, we have from (2.30) that $\varepsilon|\mathbf{x}_k - \langle \mathbf{x} \rangle| = \mathcal{O}(-\sigma_k^- + \sigma_k^+)$. Since $\lambda_{ij} = \lambda_{ji}$, we have that $\sigma_k^- = \mathcal{O}(\sigma_{k+1}^+)$. Moreover from (2.28) we have that

$|\mathbf{x}_k - \langle \mathbf{x} \rangle| = \mathcal{O}(k)$, for $\mathbf{x}_k \in \chi_k$. All this together implies that

$$\sigma_k^+ = \mathcal{O}(\sigma_k^- + \varepsilon k) = \mathcal{O}(\sigma_{k+1}^+ + \varepsilon k). \tag{2.32}$$

We now prove by induction that (2.31) holds true. We start with $k = M$ and we proceed downwards. Let $k = M$ and consider a sphere $\mathbf{x}_M \in \chi_M$, which is located at the periphery of the configuration. Consequently, all neighbours satisfy the second equation in (2.29) and we have $n_k = 1$ and $\sigma_M^- = 0$. Finally, from (2.32) we have that $\sigma_M^+ = \mathcal{O}(\varepsilon M)$. Suppose the inductive hypothesis $\sigma_{k+1}^+ = \mathcal{O}(\varepsilon(M + M - 1 + ... + k + 1))$, $\mathbf{x}_{k+1} \in \chi_{k+1}$ holds true. From (2.32) we finally get that (2.31) holds true. The maximum value for σ_k is attained at the center, where $k = 1$ and $\sigma_1 = \mathcal{O}(\varepsilon(M + ... + 1)) = \mathcal{O}(\varepsilon M^2)$. Finally, note that $\lambda_{kk_j} = \mathcal{O}(\sigma_k)$ and that since the radius of each sphere has order 1, we have from Lemma 29 that $M = \mathcal{O}(N^{1/d})$, which concludes the proof. ∎

2.4 Convergence analysis

2.4.1 Duality gap: non-convergence of AHA

We first analyze the Arrow-Hurwicz system (2.13) and more precisely the underlying saddle-point problem (2.10) for $\mathcal{L}_\varepsilon$ defined in (2.14). The goal of the Section is to show that there does not exist a saddle point which in turn explains why the Arrow-Hurwicz does not converge.

We have seen that showing existence of solution to the minimization problem (2.3), also referred to as the *primal* problem, is rather simple (prop. 25). Proving existence (or non-existence) of a saddle point for $\mathcal{L}_\varepsilon$ (2.14) is in general a more delicate manner. Let us consider the *min-max* problem:

$$d_* = \sup_{\lambda \geq 0} \inf_X \mathcal{L}_\varepsilon(X, \lambda) \tag{2.33}$$

$$p_* = \inf_X \max_{\lambda \geq 0} \mathcal{L}_\varepsilon(X, \lambda). \tag{2.34}$$

One can show that $d_* < p_*$ (weak duality). But it is more delicate to know whether $d_* = p_*$ (strong duality) in which case there exists a saddle point [27].

Remark 2.4.1. *The notations d_* and p_* are used since they each correspond to the solution of (resp.) the dual and primal problem.*

To further investigate the Lagrangian $\mathcal{L}_\varepsilon$, we introduce the functions:

$$g(\lambda) = \inf_X \mathcal{L}_\varepsilon(X, \lambda) \quad , \quad f(X) = \sup_{\lambda \geq 0} \mathcal{L}_\varepsilon(X, \lambda).$$

The solutions of the min-max problem d_* and p_* correspond (resp.) the supremum and infinimum of g and f. We first study the function f.

Lemma 31. *For any configuration X,*

$$f(X) = \begin{cases} \varepsilon W(X) & \text{if } X \in \mathcal{C} \\ +\infty & \text{if } X \notin \mathcal{C}, \end{cases} \tag{2.35}$$

where $\mathcal{C}$ (2.4) is the set of non-overlapping configuration.

Proof. If X does not satisfy the constraint (i.e. $X \notin \mathcal{C}$) then $f(X) = +\infty$. Moreover, if $X \in \mathcal{C}$ (i.e. $F(X) \leq 0$), we have $\langle \lambda, F(X) \rangle \leq 0$ and therefore $f(X) = \varepsilon W(X)$. ∎

From lemma 31, we deduce that:

$$\inf_X f(X) = \inf_{X \in \mathcal{C}} \varepsilon W(X) = \varepsilon W(\overline{X}).$$

Thus, $p_* = \varepsilon W(\overline{X})$ where $\overline{X}$ is a solution of the primal problem (2.3).

We now turn to the dual problem with the function g.

Lemma 32. *For any positive coefficients $\lambda \geq 0$,*

$$g(\lambda) = \begin{cases} 4R^2 \sum_{i<j} \lambda_{ij} & \text{if } A_\varepsilon \text{ positive definite} \\ -\infty & \text{if } A_\varepsilon \textbf{ not } \text{positive definite} \end{cases} \tag{2.36}$$

where A_ε is the matrix with coefficients:

$$a_{ij} = \begin{cases} -(\varepsilon/2N - \lambda_{ij}) & , \text{ if } i \neq j, \\ \sum_{k \neq i} (\varepsilon/2N - \lambda_{ik}) & , \text{ if } i = j. \end{cases} \tag{2.37}$$

Proof. We use the expression of the Lagrangian $\mathcal{L}_\varepsilon$ and isolate how it depends on the configuration X:

$$\mathcal{L}_\varepsilon(X, \lambda) = \sum_{i<j} \left(\frac{\varepsilon}{2N} - \lambda_{ij} \right) |\mathbf{x}_j - \mathbf{x}_i|^2 \; + \; 4R^2 \sum_{i<j} \lambda_{ij}.$$

Thus, we can write the Lagrangian as a quadratic form in X:

$$\mathcal{L}_\varepsilon(X, \lambda) = \langle A_\varepsilon X, X \rangle \; + \; 4R^2 \sum_{i<j} \lambda_{ij}, \tag{2.38}$$

where the matrix A_ε is given by (2.37). Indeed, for any symmetric coefficients c_{ij}, denoting

$\sigma_i = \sum_j c_{ij}$, we have:

$$\begin{aligned}\sum_{i<j} c_{ij}|\mathbf{x}_j - \mathbf{x}_i|^2 &= \frac{1}{2}\sum_{i,j} c_{ij}(\mathbf{x}_j^2 + \mathbf{x}_i^2 - 2\mathbf{x}_j\mathbf{x}_i) \\ &= \sum_i \sigma_i \mathbf{x}_i^2 - \sum_{i,j} c_{ij}\mathbf{x}_i\mathbf{x}_j = MX,\end{aligned}$$

where M is a matrix with entries $-c_{ij}$ outside the diagonal and $\sum_{k\neq i} c_{ik}$ on the diagonal.

Using the formulation (2.38), we deduce that the minimization of $\mathcal{L}_\varepsilon$ is $-\infty$ if the matrix A_ε (2.37) is **not** positive definite. On the other hand, if A_ε is positive definite (i.e. $\langle A_\varepsilon X, X\rangle \geq 0$) then (2.38) leads to:

$$\mathcal{L}_\varepsilon(X,\lambda) \geq 4R^2 \sum_{i<j} \lambda_{ij}.$$

Therefore the infinimum is reached with $X = 0$ and $g(\lambda) = 4R^2 \sum_{i<j} \lambda_{ij}$. ∎

It is delicate to find explicit condition for the matrix A_ε to be positive definite. One can find sufficient condition: if $(\varepsilon/2N - \lambda_{ij}) \geq 0$ then A_ε is diagonal dominant and therefore positive definite. Moreover, A_ε is **not** positive definite if the diagonal is not positive. Thus, let $\sigma_i = \sum_k \lambda_{ik},\ i = 1, ..., N$. If there exists i such that $\sigma_i > \varepsilon(N-1)/2N$, then the matrix A_ε is **not** positive definite.

Proposition 33. *The solution of the dual problem is given by:*

$$d_* = \sup_{\lambda\geq 0} g(\lambda) = \varepsilon R^2(N-1). \tag{2.39}$$

Proof. Denoting again $\sigma_i = \sum_k \lambda_{ik}$, the matrix A_ε (2.37) is not positive definite if there exists $\sigma_i > \varepsilon(N-1)/2N$ and therefore $g(\lambda) = -\infty$ in this case. We deduce that the maximization of the function g can be reduced over the λ satisfying $\sigma_i \leq \varepsilon\frac{N-1}{2N}$ for all i.

From lemma 32, we have:

$$g(\lambda) \leq 4R^2 \sum_{i<j} \lambda_{ij} = 2R^2 \sum_i \sigma_i.$$

Therefore, $g(\lambda) \leq \varepsilon R^2(N-1)$. This upper-bound is reached by taking $\lambda_{ij} = \varepsilon/2N$ for $i \neq j$ leading to the result (2.39). ∎

Notice that the value d_* corresponds to configurations where all spheres are at a distance $2R$ from each other. Indeed, suppose $|\mathbf{x}_j - \mathbf{x}_i| = 2R$ for all $i \neq j$, then:

$$W(X) = \frac{\varepsilon}{2N}\sum_{i<j} 4R^2 = \frac{N(N-1)}{4N} 4\varepsilon R^2 = (N-1)\varepsilon R^2.$$

Thus, we have $d_* = p_*$ **only** if all the spheres can be at distance $2R$ from each other. In dimension 2, this corresponds to having only $N=2$ or $N=3$ spheres. But if $N \geq 4$ in dimension 2, then $d_* < p_*$ and therefore there is no saddle point (X_*, λ_*) for the Lagrangian $\mathcal{L}$ in this case.

Theorem 34. *The Lagrangian $\mathcal{L}_\varepsilon$ (2.14) does* **not** *have a saddle point if $N > d+1$ where N is the number of spheres and $d \leq 4$ is the spatial dimension.*

While we do not have a saddle point for the Lagrangian (2.9), we can still look for *local saddle* points. In short, does it exist a point $(\widetilde{X}, \widetilde{\lambda})$ satisfying:

$$\mathcal{L}_\varepsilon(\widetilde{X}, \lambda) \leq \mathcal{L}_\varepsilon(\widetilde{X}, \widetilde{\lambda}) \leq \mathcal{L}_\varepsilon(X, \widetilde{\lambda}) \qquad \text{for all } (X, \lambda) \textit{ close to } (\widetilde{X}, \widetilde{\lambda}). \tag{2.40}$$

The hope is that solutions of the primal problem $\overline{X}$ (with the corresponding Lagrangian multiplier $\overline{\lambda}$) will be local saddle point (i.e. solution of (2.40)). Then, we could apply the Arrow-Hurwicz algorithm to find those solutions.

The left inequality of (2.40), i.e. $\mathcal{L}_\varepsilon(\widetilde{X}, \lambda) \leq \mathcal{L}_\varepsilon(\widetilde{X}, \widetilde{\lambda})$, imposes that $\widetilde{X}$ satisfies the constraints. The right inequality means that $\widetilde{X}$ is a local minimum of:

$$X \mapsto \mathcal{L}_\varepsilon(X, \widetilde{\lambda}). \tag{2.41}$$

Using the formulation (2.38), we deduce that $\widetilde{X}$ is a local minimum if the matrix A_ε (2.37) is positive definite. As we have seen previously, in order to have the matrix A_ε positive definite, it is necessary to have the coefficients λ_{ij} satisfying: $\sigma_i = \sum_k \lambda_{ik} \leq \varepsilon \frac{N-1}{2N}$. This property is not satisfied by the Lagrange multiplier. They are actually of the order of (at least) $\varepsilon N^{1/d}$ (see Proposition 30). We conclude that a solution to the primal problem $(\overline{X}, \overline{\lambda})$ cannot be a local saddle point, as it is not a local minimum to the function defined in (2.41).

2.4.2 Stability analysis in $d = 1$: convergence of DAHA and ALA

We now analyse the ODE systems (2.16) and (2.19) associated to the ALA and DAHA, respectively, for an arbitrary number of spheres N in dimension $d = 1$. Note that such systems are piecewise smooth. In particular, they are smooth in a neighbourhood of $\overline{X}$, which allows us to carry out the linear stability analysis in order to study the local convergence of the solution towards a steady state. We will show that the steady state obtained in Section 2.3.1 is an asymptotically stable steady state of both systems as defined in Chapter 1, Definition 14. Note that this definition assumes that the initial configuration is chosen close enough to the steady state. The next Theorem gives conditions on the function f that ensure the asymptotic stability of a given steady state of the ODE system $\dot{x} = f(x),\ t \geq 0$.

Theorem 35. *Let $n \in \mathbb{N}$ and let $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ be a function. Consider the ODE system $\dot{x} = f(x)$ and a steady state x^*, where f is smooth at x^*. If the eigenvalues of the Jacobian matrix of f evaluated at x^*, $Df(x^*)$, have negative real part, then x^* is asymptotically stable. Moreover, let ϕ_t be the flow of the differential equation. Then if $-\xi$ is a number larger than every real part of an eigenvalue of $Df(x^*)$, then there is a neighbourhood U of x^* and a constant $k > 0$ such that*

$$|\phi_t(x) - x^*| \leq k|x|e^{-\xi t},$$

whenever $x \in U$ and $t \geq 0$.

Proof. See [34], thm. 2.42, p. 158 and cor. 2.43, p. 160. ∎

Next we prove that a solution to the minimization problem (2.3) corresponds to an asymptotically stable steady state of the ODE systems (2.16) and (2.19) in the case of an arbitrary number N of spheres in $\mathbb{R}$.

Proposition 36. *Let $\overline{X} = (\overline{x}_i)_{i=1,...,N}$ be a solution to the minimization problem* (2.3) *and $\overline{\lambda} = (\overline{\lambda}_{ij})_{i<j}$ the associated Lagrange multipliers in dimension $d = 1$. Then, for any ε, γ positive, $(\overline{X}, \overline{\lambda})$ is an asymptotically stable steady state of both systems ALS* (2.16) *and DAHS* (2.19)*.*

Proof. First note that by adding up all differential equations in $x_i,\ i = 1, ..., N$ of (2.16) we get that $\langle \dot{x} \rangle = 0$, ie., the center of mass is conserved. So without loss of generality, let $\langle x \rangle = 0$. From Prop. 27, the solution to the minimization problem is unique (apart from permutations) and satisfies $\overline{x}_{i+1} - \overline{x}_i = 2R$, $\overline{\lambda}_{i,i+1} > 0$ and

$$\phi_{ij}(\overline{X}) < 0,\ \overline{\lambda}_{ij} = 0,\ j \neq i-1, i+1 \tag{2.42}$$

Second, from (2.15) and (2.6) it follows that $(\overline{X}, \overline{\lambda})$ is a critical-point of the Lagrangian (2.14), and therefore it is also a steady state of (2.16). We want to prove that this steady state is asymptotically stable. Consider a close enough perturbation $(X(0), \lambda(0))$ of $(\overline{X}, \overline{\lambda})$ satisfying

$$|x_i(0) - \overline{x}_i| < R,\ i = 1, ..., N,$$

$$\lambda_{i,i+1}(0) \geq 0,\ i = 1, ..., N-1 \text{ and } \lambda_{ij}(0) = \overline{\lambda}_{ij} = 0,\ j \neq i+1, i-1,$$

$$\lambda_{i,i+1}(0) + \gamma\phi_{i,i+1}(X(0)) > 0,\ i = 1, ..., N-1$$

Note that we only perturb the Lagrange multipliers that are associated to the active constraints at $\overline{\lambda}$. Indeed, from (2.42) follows that $\phi_{ij}(X(t)) < 0, j \neq i+1, i-1$ for t in some time interval $[0, T]$, consequently, $\dot{\lambda}_{ij}(t) < 0$ and any small enough perturbation would decay to zero.

Consider the following smooth system:

$$\begin{cases} \dot{x}_1 &= -\varepsilon x_1 + 2(\lambda_{12} + \gamma\phi_{12})(x_1 - x_2) \\ \dot{x}_i &= -\varepsilon x_i + 2(\lambda_{i,i-1} + \gamma\phi_{i,i-1})(x_i - x_{i-1}) + 2(\lambda_{i,i+1} + \gamma\phi_{i,i+1})(x_i - x_{i+1}), \\ & \qquad\qquad\qquad\qquad\qquad\qquad i = 2, ..., N-1 \\ \dot{x}_N &= -\varepsilon x_N + 2(\lambda_{N-1,N} + \gamma\phi_{N-1,N})(x_N - x_{N-1}) \\ \dot{\lambda}_{i,i+1} &= \phi_{i,i+1}, \quad i = 1, ..., N-1 \\ \lambda_{ij} &= 0, \ j \neq i-1, i+1 \end{cases} \tag{2.43}$$

Now, we note two aspects. First, if

$$|x_i(t) - \overline{x}_i| < R, \ i = 1, ..., N, \ t \geq 0 \tag{2.44}$$

holds true, meaning that no two spheres cross each other at any moment, then for $j \neq i-1, i+1$, $\phi_{ij}(x(t)) < 0$ and $\lambda_{ij}(t) = 0$ for all $t \geq 0$ and therefore $\dot{\lambda}_{ij}(t) = 0, \ j \neq i-1, i+1$ for all $t \geq 0$. Consequently, the system (2.16) is equivalent to (2.43). Second, if $(\overline{X}, \overline{\lambda})$ is an asymptotically stable steady state of (2.43), then condition (2.44) holds for all times, provided that the initial condition $(X(0), \lambda(0))$ is chosen close enough to the steady state.

It is then enough to prove that $(\overline{X}, \overline{\lambda})$ is an asymptotically stable steady state of (2.43).

We first compute the difference $\dot{x}_{i+1} - \dot{x}_i, \ i = 1, ..., N-1$. Considering the change of variables $x_{i+1,i} := x_{i+1} - x_i$ we get

$$\begin{aligned} \dot{x}_{i+1,i} &= -\varepsilon x_{i+1,i} + 4(\lambda_{i+1,i} + \gamma\phi_{i+1,i})x_{i+1,i} \\ & \quad -2(\lambda_{i+1,i+2} + \gamma\phi_{i+1,i+2})x_{i+2,i+1} - 2(\lambda_{i,i-1} + \gamma\phi_{i,i-1})x_{i,i-1}. \end{aligned} \tag{2.45}$$

We will now write the system describing the dynamics of the $2(N-1)$ variables $x_{i,i+1}, \ \lambda_{i,i+1}, \ i = 1, ..., N-1$ in matrix form. Note that these are all the variables of system (2.43), because as we saw $\lambda_{ij} = 0, \ i \neq j+1, j-1$ and $\langle x \rangle = 0$ for all $t \geq 0$. Consider the $(N-1) \times (N-1)$ square matrices X, Λ and A defined by

$$X = \begin{bmatrix} x_{21} & & & 0 \\ & x_{32} & & \\ & & \ddots & \\ 0 & & & x_{N,N-1} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_{21} & & & 0 \\ & \lambda_{32} & & \\ & & \ddots & \\ 0 & & & \lambda_{N,N-1} \end{bmatrix}$$

$$\text{and} \qquad A = \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & & -1 \\ 0 & & & -1 & 2 \end{bmatrix}$$

and the $(N-1)$ vectors $\lambda = [\lambda_{i+1,i}]_{i=1,\dots,N-1}$ and $\mathbb{1} = [1]_{i=1,\dots,N-1}$. Then, equations (2.45) can be written in matrix form as

$$\dot{X}\mathbb{1} = -\varepsilon X\mathbb{1} + 2AX\lambda + 2\gamma 4R^2 AX\mathbb{1} - 2\gamma AX^3\mathbb{1} \tag{2.46}$$

Using the same notation, we may also write the equation for λ in matrix form as follows

$$\dot{\lambda} = (4R^2 I - X^2)\mathbb{1}, \tag{2.47}$$

where I is the $(N-1)\times(N-1)$ identity matrix. The dynamics of the system (2.43) can then be described by

$$\begin{cases} \dot{X}\mathbb{1} &= -\varepsilon X\mathbb{1} + 2AX\lambda + 2\gamma 4R^2 AX\mathbb{1} - 2\gamma AX^3\mathbb{1} &:= F(X,\lambda) \\ \dot{\lambda} &= (4R^2 I - X^2)\mathbb{1} &:= G(X,\lambda) \end{cases}$$

After a simple computation we conclude that the steady states of this system satisfy, $X^{*2} = 4R^2 I$ and $\lambda^* = \frac{\varepsilon}{2} A^{-1}\mathbb{1}$. In particular,

$$(\overline{X}, \overline{\lambda}) = (2RI, \frac{\varepsilon}{2} A^{-1}\mathbb{1}) \tag{2.48}$$

is a steady state, as expected. We want to find conditions on the parameters ε and γ such that $(\overline{X}, \overline{\lambda})$ is asymptotically stable. The Jacobians of F and G are given by

$$D_{(X,\lambda)}(F,G) = \begin{bmatrix} D_X F & D_\lambda F \\ D_X G & D_\lambda G \end{bmatrix}$$

with

$$\begin{aligned} D_X F(X,\lambda) &= -\varepsilon I + 2A\Lambda + 2\gamma 4R^2 A - 2\times 3\gamma AX^2 \\ D_\lambda F(X,\lambda) &= 2AX \\ D_X G(X,\lambda) &= -2X \\ D_\lambda G(X,\lambda) &= 0 \end{aligned}$$

The Jacobian evaluated at the steady state (2.48) is given by the block matrix

$$D_{(X,\lambda)}(F,G)|_{(\overline{X},\overline{\lambda})} = 4R \begin{bmatrix} -4R\gamma A & A \\ -I & 0 \end{bmatrix} \tag{2.49}$$

Since this matrix is negative definite for all γ positive, then by Theorem 35 from Section 0.3.2.1 the steady state (2.48) is asymptotically stable.

Applying the same reasoning to the damped Arrow-Hurwicz system (2.19) we obtain the Jacobian evaluated at the steady state (2.48),

$$D_{(X,\lambda)}(F,G)|_{(\overline{X},\overline{\lambda})} = 4R \begin{bmatrix} -2R\gamma\varepsilon I & A \\ -I & 0, \end{bmatrix} \tag{2.50}$$

which is positive definite for all γ and ε positive and again by Theorem 35 the steady state (2.48) is also asymptotically stable. This concludes the proof. ∎

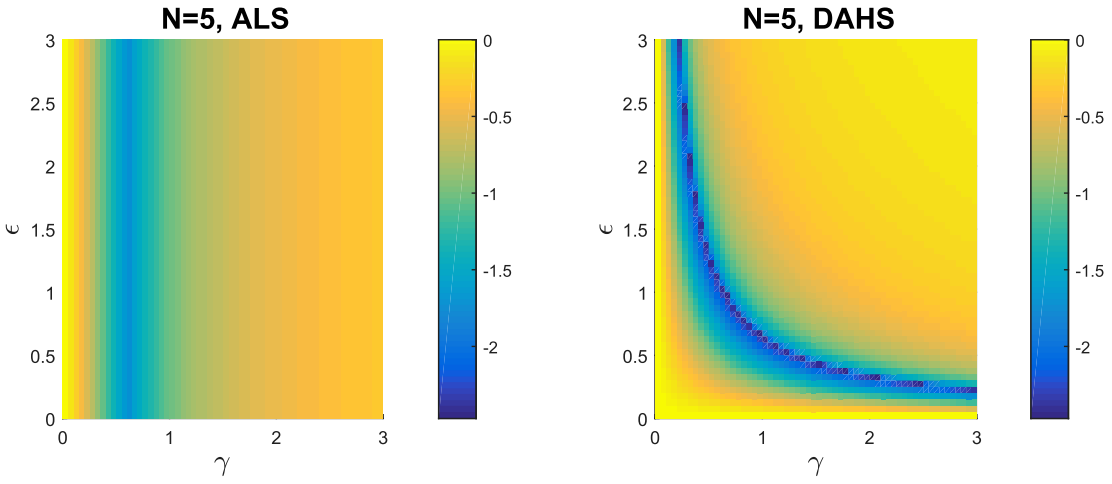


Figure 2.4: Largest real part of the eigenvalues of the Jacobian matrices (2.49) and (2.50) for the Augmented Lagrangian system (left) and for the damped Arrow-Hurwicz system (right), respectively, as a function of the parameters (γ, ε) for $N = 5$ and $R = 1$. We observe that the eigenvalues can be smaller in the case of the DAHS, but the region of small eigenvalues is narrower when compared to the case of ALS.

This proposition allows us to conclude that the DAHS and the ALS both converge to a solution of the minimization problem in dimension $d = 1$ for any positive value of the parameters ε and γ, provided the initial configuration is close enough to the steady state. Theorem 35 also tells us that the speed of convergence is exponential with a rate of decay that is larger than the largest eigenvalue of the Jacobian matrix (2.49), in the case of ALS, and (2.50) in the case of the DAHS. We plot in Figure 2.4 the largest real part of the eigenvalues of each of these matrices as a function of the parameters $(\gamma, \varepsilon) \in [0, 3]^2$ for $N = 5$. In Figure 2.5 we plot the same as in Figure 2.4 but for $N = 100$ and for the range of parameters $(\gamma, \varepsilon) \in [0, 30]^2$ in the case of the AL and $(\gamma, \varepsilon) \in [0, 3]^2$ in the case of the DAHS. In both Figures, we can see that the maximum eigenvalue is always negative, which is expected. The blue region corresponds to the parameter region where

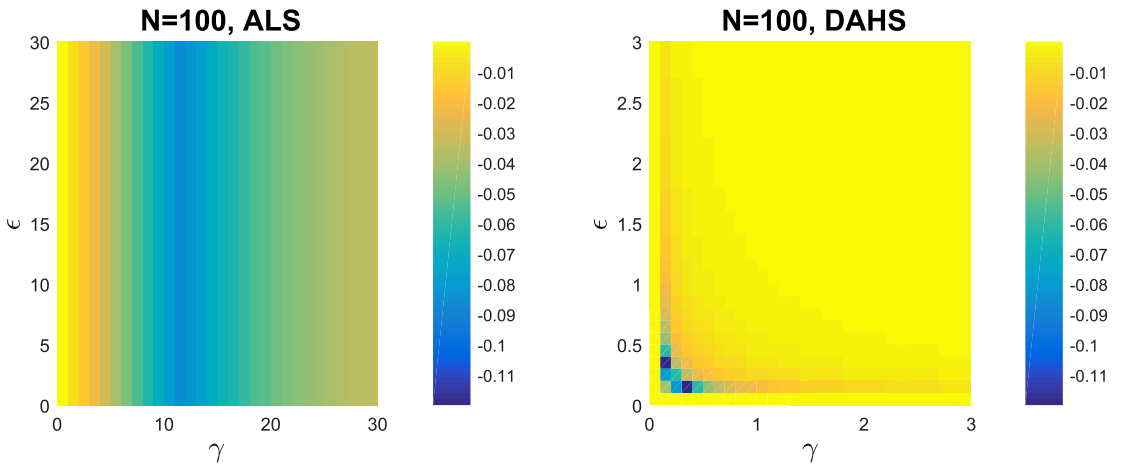


Figure 2.5: The same as in Figure 2.4 for the case of $N = 100$. The conclusions are similar to the case of $N = 5$ presented in Figure 2.4.

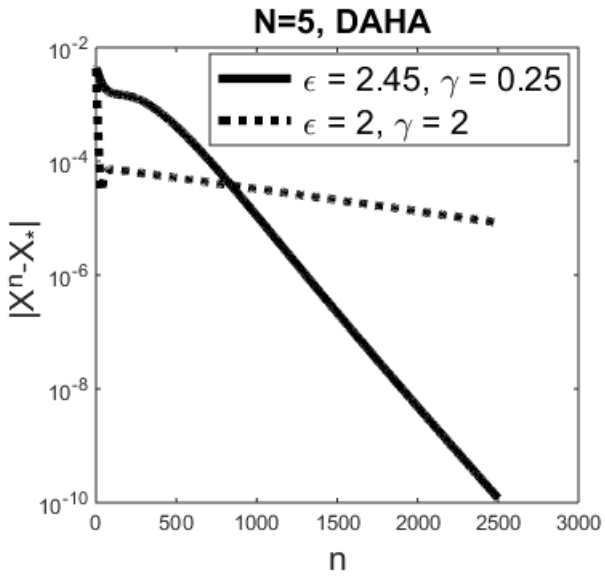


Figure 2.6: Euclidean distance between the solution obtained numerically at each iteration by the DAHA and the steady state given in Proposition 27 X_*, i.e., $|X^n - X_*|$ for two different sets of parameters taken from the yellow (dot line) and blue (solid line) regions of the graph in Figure 2.4-right. As expected, the solid line decays at a larger rate than the dashed line. The model parameters are $N = 5$, $R = 1$ and the numerical parameter $\Delta t = 0.002$.

the maximum eigenvalue is smaller, which may lead to a faster convergence towards equilibrium. Comparing the two systems, the blue region in the case of the DAHS is darker, meaning that the eigenvalues are smaller, than in the case of ALS. On the other hand, the blue region is wider in the case of the ALS compared to the DAHS. This indicates that the DAHS may be faster for specific parameters, but the region where to choose the best parameters is narrower than for the ALS. Comparing the results for each value of N, we observe that for $N = 100$ the eigenvalues are one order of magnitude larger compared to the case with $N = 5$. Moreover, for the DAHS the region with small eigenvalues has became narrower for N large, while for the ALS the corresponding region became wider.

The Euclidean distance between a numerical solution obtained with the DAHA and the equilibrium state given in Proposition 27 over iteration number is plotted in Figure 2.6 for $N = 5$ and for two different pairs (ε, γ) taken from the yellow and blue region of Figure 2.4, respectively. We see that starting from a distance of the order of 10^{-2}, the convergence is exponential, and the system converges much faster (decay rate of approximately 0.008) for values of the parameters chosen from the blue region than the yellow one (decay rate of approximately 0.001).

2.4.3 Numerical simulations in dimension $d = 1$

In this Section we present simulations of the evolution of the dynamical systems AHS (2.13) and ALS (2.16) in dimension $d = 1$ in order to illustrate the analytical results presented in the previous Sections. The conclusions regarding the ALS also apply to the DAHS, so we do not consider this method in this Section.

To illustrate the non-convergence of the AHS, we present in Figure 2.7 the evolution of the positions X with $N = 3$ particles, which corresponds to the simplest scenario where the Lagrangian $\mathcal{L}$ (and $\mathcal{L}_\varepsilon$) does not have a saddle point thanks to Theorem 34. We observe that the trajectories $(x_1(t), x_2(t), x_3(t))$ have a chaotic behavior for $\varepsilon = .5$ (left Figure) whereas it converges to a periodic solution with $\varepsilon = 1$ (right Figure). The convergence of the AHS to a periodic solution occurs in three steps. Phase I: the particles move away from each other (i.e. λ_{ij} and $\mathcal{L}$ increases) until there is no more overlapping (i.e. $X(t)$ moves to the constraint set $\mathcal{C}$). Phase II: the solution contracts (i.e. $X(t)$ moves back to the border of $\mathcal{C}$). Phase III: the solution starts to oscillate. For instance, in Figure 2.7-right, phase I corresponds roughly to $t \in [0, 1]$, phase II: $t \in [1, 3]$ and phase III: $t \geq 3$.

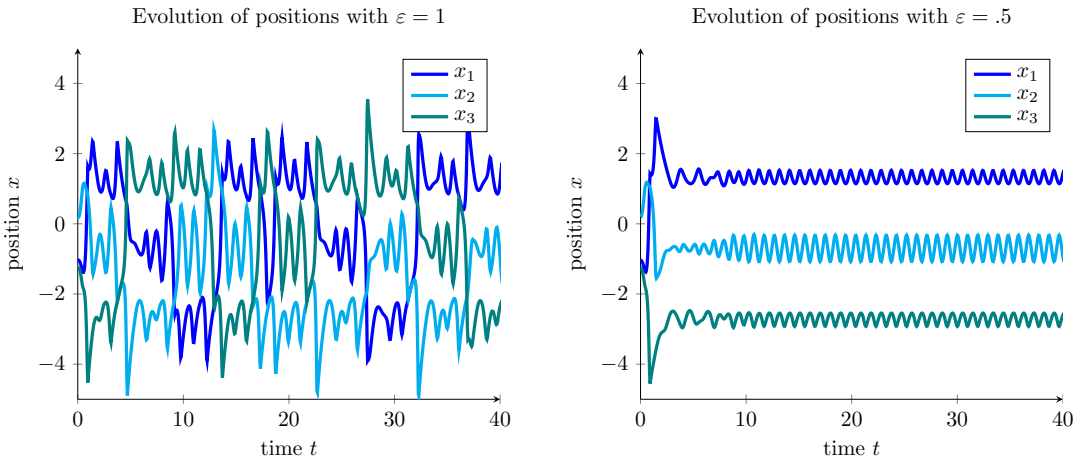


Figure 2.7: Solution to the AHS (2.13) with $N = 3$ particles in one dimension with $\varepsilon = 1$ (left Figure) and $\varepsilon = 0.5$ (right Figure). A period solution occurs only if ε is small enough. Parameters: $\Delta t = 0.05$, $R = 1$.

A similar situation has been observed when we increased the number of particles N. The dynamical system (2.13) periodic solution only if ε is small enough.

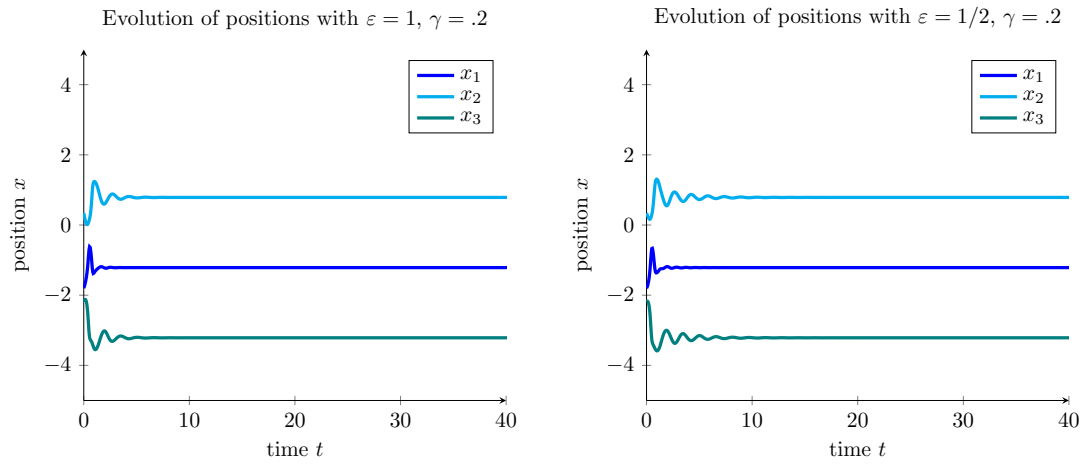


Figure 2.8: Similar simulations as in Figure 2.7 using the stabilizer dynamics ALS (2.16) with $\gamma = 0.2$. Both simulations converge to a steady state.

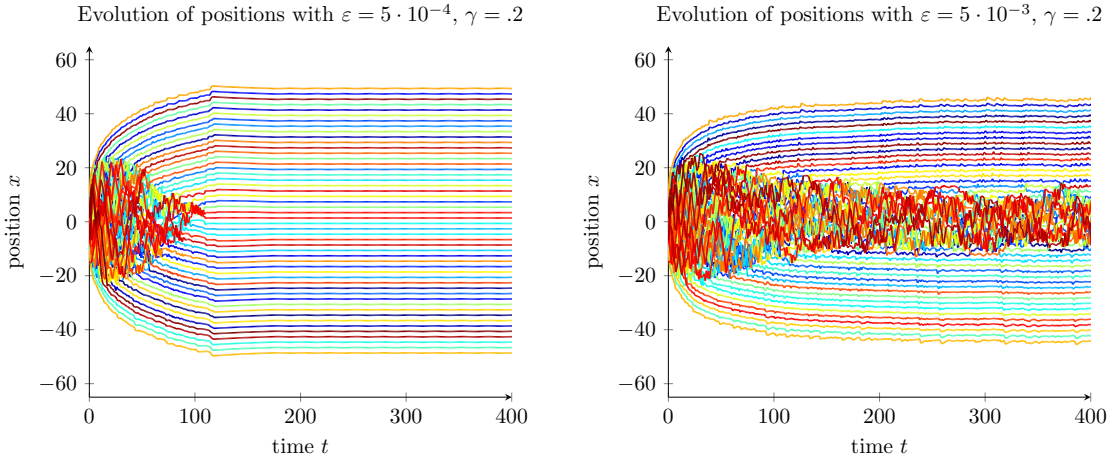


Figure 2.9: Simulation with ALS (2.16) for $N = 50$ particles and for $(\varepsilon, \gamma) = (5 \cdot 10^{-4}, 0.2)$ (left) and $(\varepsilon, \gamma) = (5 \cdot 10^{-3}, 0.2)$ (right). The dynamics converges if both ε is 'small enough' and γ large enough. Parameters: $\Delta t = 0.05$, $R = 1$.

We now run the same simulations but with the augmented Lagrangian system (2.16). We observe in Figure 2.8 that using $\gamma = .2$ the solution converges to a steady state. In the simulation with $N = 50$ particles in Figure 2.9-left, adding the stabilizer makes the dynamics converge to a steady state. However, if ε is too large as in Figure 2.9-right, the dynamics is chaotic at the center and it does not seem to converge. This indicates that despite being an asymptotically stable steady state of the modified system, if the initial configuration is far away from the steady state, then the solution may not converge to it. This situation should be further explored in the future. In particular, given an arbitrary initial configuration, what parameter regime leads to the convergence of the ALA towards the steady state.

2.5 Conclusion and future work

In this Chapter we study a non-convex minimization problem with non-overlapping constraints that naturally arises in the study of particle systems. We show that the Arrow-Hurwicz algorithm does not converge and we prove convergence of the augmented Lagrangian algorithm and of the damped Arrow-Hurwicz algorithm (DAHA) towards a solution of the minimization problem in dimension $d = 1$.

There are still many questions that remain unanswered and could be investigated in the future. Two of them are concerned with the convergence properties of the ALA and DAHA in the case of higher spacial dimensions and how we should choose the parameters to get the fastest convergence. One could also investigate the types of local minima that may arise from different types of initial conditions and the convergence properties of the more general 2nd-order DAHA developed in Chapter 1.

Despite of the simplicity of the formulation and resolution of sphere packing problems in low dimensions, they turn out to be extremely complex in higher dimensions (above 2). Therefore they still remain nowadays an active area of research in mathematics [36, 89].

Part II

Particle-based models for congested systems

3

Event-driven versus time-stepping schemes for ballistic aggregation

This Chapter deals with the problem of simulating the dynamics of large and densely packed particle systems subject to ballistic aggregation. Two different paradigms are typically used: event-driven and time-stepping algorithms. Despite being more accurate, event-driven algorithms become computationally very expensive as the number of particles increases. Such situations require the use of an alternative approach, such as time-stepping algorithms. In these methods, the contact events occurring during a small time-interval are grouped and solved simultaneously. In this work we explore a new time-stepping approach for the case of ballistic aggregation and we compare it to the event-driven one in terms of computational time. This approach is based on a minimization problem with non-overlapping constraints that is solved by the damped Arrow-Hurwicz algorithm developed in Part I, which distinguishes it from other algorithms proposed in the literature. Specifically, we obtain four time-stepping algorithms from the combination of considering smooth or non-smooth constraints and rigid or non-rigid aggregates. Numerical results obtained in the case of a bounded domain and spherical particles show that when the number of particles becomes large or the size of the domain becomes small the time-stepping methods perform faster than the event-driven. The threshold at which this switch occurs is identified. In particular, for $N = 10^6$ particles, the event-driven takes almost four days, while the time-stepping takes only a few hours. Among the four time-stepping schemes, the ones with smooth constraints seem to perform faster for any number of particles.

This work is based on a collaboration with Giacomo Dimarco[*]*, Sophie Hecht and Pierre Degond*[†]*.*

[*]Department of Mathematics and Computer Science, University of Ferrara, Italy
[†]Department of Mathematics, Imperial College London, UK

3.1 Introduction

Aggregation can be found in many different fields of physics, such as aerosols, raindrops formation, sprays, polymers, formation of planets and galaxies, nucleation and growth, crystallization [6, 175, 56]. Also in biology, cells aggregate to form tissues or to repair injuries (blood coagulation), bacteria aggregate to form biofilms [106], etc. Aggregation can be observed and studied at different scales [58]. For instance, at the microscopic level one gets particle based models describing the evolution of the mass, position and velocity of each particle [69, 70, 116, 48, 119, 120, 169, 142, 79]. At the mesoscopic level, one studies the evolution of a statistical distribution of sizes, positions and velocities of the aggregates [154, 58, 94]. Within this description, a widely studied approach is given by the so-called Smoluchowski coagulation equations or the Marcus-Lushnikov model [6, 67]. At the macroscopic level, one typically describes averages of the distribution function with respect to the mass or size of the clusters [113, 57]. However, despite of being easier to treat, both computationally and analytically, the mesoscopic and macroscopic models do not incorporate all information about position or velocity of the clusters [6]. Moreover, it is often assumed that the clusters are spherical. A particle based model instead is able to give more detailed spatial information and it allows the construction of more realistic models [50, 79]. Additionally, it provides an appropriate framework to study the randomness arising on the initial conditions [70, 116], the statistical properties of the growing aggregates related to the evolution of the shape [99, 128, 13] and the type of aggregates obtained in the long-time limit [119]. For these reasons we opt by the microscopic approach.

Specifically, we focus on a particular type of aggregation phenomenon, called ballistic aggregation, in which hard-spheres move in straight trajectories until they collide, stick and form growing moving aggregates [56, 175]. We consider that the collisions are inelastic and frictionless and that during a collision the total mass and momentum is conserved while kinetic energy is dissipated. This model is among the simplest models for the interaction of rigid particles, which contributes for its popularity [144]. The choice of this model is justified by the fact that, despite of its simplicity, it already presents many of the difficulties related to the simulation of densely packed systems, which makes it a good system for testing and comparing numerical methods.

There are two main classes of numerical methods to simulate the dynamics of particle based models with contact events. The first is the so-called event-driven method (ED) [69, 70]. In this approach, the time of the next collision is computed, the system evolves to that time and the collision is performed exactly. The main difficulty is to determine the sequence of particle collisions. In order to simplify the dynamics, some approaches use fixed lattices space discretizations [48, 70, 119, 169, 120, 175, 63, 156], however such a simplification produces effects on the dynamics and on the shapes of the clusters [75]. The second class of methods is the so-called time-driven or time-stepping (TS) method [81].

In this approach, the time is discretized in equal intervals and all events occurring within a time-interval are solved simultaneously. The procedure at each time-step consists of computing the positions and velocities of the particles from their own individual dynamics and correcting possible overlaps between particles by solving a minimization problem with non-overlapping constraints. To solve this minimization problem we use the first-order damped Arrow-Hurwicz algorithm [45] developed in Part I. We consider four time-stepping methods which result from the combination of two features. The first feature is related to the formulation of the non-overlapping constraints, namely we consider a smooth and a non-smooth form that lead to equivalent constraints. The second feature is related with a modelling assumption on the rigidity of the clusters: we consider rigid clusters, in which the particles in the same cluster are subjected to the same dynamics and consequently, they remain always in the same relative positions, and non-rigid clusters, where the particles of each cluster move individually and adhesive particles may even fall apart within the cluster.

Since it is an exact method, an ED algorithm should be more accurate at describing the real dynamics of the system than a TS scheme. Nevertheless, ED algorithms may become computationally expensive as the number of particles increases. The aim of this work is to compare TS and ED algorithms in terms of computational time and identify in which situations it is necessary to use a TS scheme due to the high computational time an ED requires. In particular we study the influence of the number of particles and of the size of the domain. We observe that there is a threshold on these two quantities at which the TS method becomes more efficient than the ED. The numerical experiments have been run up to $N = 10^6$ particles. For this value of N, the TS algorithms took a few hours, while the ED took a few days to simulate the dynamics.

The plan of the Chapter is the following. In Section 3.2 we introduce the time-continuous model for ballistic aggregation and the ED method. In Section 3.3 we present a corresponding time-discrete dynamics and the four TS schemes. In Section 3.4, we first optimize the parameters involved in the TS schemes. Then we compare the ED and the TS methods in terms of computational speed for different values of N and size of the domain. In the final Section 3.5, we draw some conclusions and suggest future investigations.

3.2 Time-continuous model and event-driven (ED) scheme

3.2.1 Time-continuous model

Let $d, N \in \mathbb{N}$ and $L > 0$. We consider a system of N self-propelled hard spheres on a $d-$dimensional torus $[0, L]^d$. The spheres are characterized by their radius $R_i > 0$, their position $X_i = (x_{i_1}, ..., x_{i_d})$, their velocity $V_i = (v_{i_1}, ..., v_{i_d})$ and their mass $m_i >$

$0,\ i = 1, ..., N$. They interact via inelastic frictionless collisions and an adhesive force is immediately created between the colliding particles. No other forces are exerted on the spheres. Thus, between two consecutive collisions, particles travel in straight trajectories at a constant speed. Given the times t_1 and t_2 of two consecutive collisions, the velocities $\mathbf{V}(t) = \{V_i(t)\}_{i=1,...,N}$ and positions $\mathbf{X}(t) = \{X_i(t)\}_{i=1,...,N}$ of the particles at time $t \in [t_1, t_2]$ satisfy the following equations on the torus

$$\frac{dX_i}{dt}(t) = V_i(t_1),\ t \in [t_1, t_2],\ i = 1, ..., N. \tag{3.1}$$

For each $i, j = 1, ..., N,\ i < j$, consider the functions $\phi_{ij} : \mathbb{R}^{dN} \to \mathbb{R},\ i < j$, defined by,

$$\phi_{ij}(\mathbf{X}) = R_i + R_j - |X_i - X_j|, \tag{3.2}$$

where $|.|$ denotes the Euclidean norm on the torus (see Remark 3.2.1).

Remark 3.2.1. *The Euclidean norm of a vector $X = \{x_i\}_{i=1,...,d}$ on a torus $[0, L]^d$ is defined by*

$$|X|^2 = \min_{\mathcal{P}_1, \mathcal{P}_2:\ \mathcal{P}_1 \cup \mathcal{P}_2 = \{1,...,N\},\ \mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset} \left\{ \sum_{i \in \mathcal{P}_1} |x_i|^2 + \sum_{i \in \mathcal{P}_2} (L - |x_i|)^2 \right\}, \tag{3.3}$$

The norm induces the Euclidean distance on a torus $d(X, Y) = |X - Y|,\ X, Y \in [0, L]^d$. This distance corresponds to select the shorter path among all possible straight paths between X and Y. An example of the four possible straight paths between two points in $\mathbb{R}^2$ is presented in Figure 3.1. In the middle of the domain the usual Euclidean distance in $\mathbb{R}^d$ holds. In order to simplify the presentation of the algorithms, we assume in the following that we are in $\mathbb{R}^d$.

In order to avoid overlapping spheres, we impose non-overlapping constraints,

$$\phi_{ij}(\mathbf{X}) \leq 0, i, j = 1, ..., N,\ i < j. \tag{3.4}$$

The collisions between particles produce discontinuities in the velocity field due to the non-overlapping constraint imposed in the system. The collision rule is described next in dimension $d = 2$. Notice that it is possible to generalize this rule to higher dimensions [171], however for simplicity we only present the two-dimensional case here. The physical background of what follows can be found for example in [16].

Let $d = 2$. When spheres i and j collide they bind and form a new moving cluster. Let $X_i,\ X_j$ be the positions of the spheres at collision and $V_i,\ V_j$ the pre-collision velocities. The center of mass of the new cluster is defined by $X_c := (m_i X_i + m_j X_j)/(m_i + m_j)$. The velocity V_c' of the center of mass of the new cluster is given by the weighted average

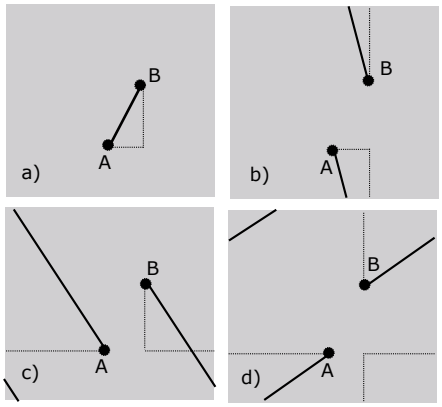


Figure 3.1: The four different straight lines going from point A to point B in a 2-dimensional torus $[0, L]^2$. The squared Euclidean norm of the vector $A - B = (x_1, x_2) \in [0, L]^2$ corresponds to the minimum between the length of each line, namely a) $x_1^2 + x_2^2$, b) $x_1^2 + (|x_2| - L)^2$, c) $(|x_1| - L)^2 + x_2^2$ and d) $(|x_1| - L)^2 + (|x_2| - L)^2$.

between the pre-collision velocities, i.e.,

$$V_c' = \frac{m_i V_i + m_j V_j}{m_i + m_j}.$$

The cluster starts to spin with an angular velocity Ω' given by

$$\Omega' = \frac{M_c}{J}$$

with

$$M_c = m_i X_i \times V_i + m_j X_j \times V_j - (m_i + m_j) X_c \times V_c',$$

where $\times$ denotes the cross product, and J is the moment of inertia given by

$$J = |X_i - X_c|^2 m_i + |X_j - X_c|^2 m_j.$$

The velocity of each sphere after the collision, V_i', V_j', is given by the sum between the velocity of the cluster and the relative velocity of the sphere with respect to the center of mass of the cluster, i.e.,

$$V_k' = V_c' + \Omega' \times (X_k - X_c), \ k = i, j.$$

Since there are no other forces acting on the particles, then the colliding particles remain near each other ever since after the collision. During the time evolution of the system, collisions occur and growing aggregates of multiple spheres are formed. In this situation, when a cluster C_1 collides with a cluster C_2 they stick together forming a new cluster C' with center of mass $X_c := \sum_{i \in C'} m_i X_i / (\sum_{i \in C'} m_i)$. The linear velocity V_c' of the center of mass of the new cluster is a weighted average between the pre-collision velocities V_i of

the spheres that constitute the new cluster,

$$V_c' = \frac{\sum_{i\in C'} m_i V_i}{\sum_{i\in C'} m_i}. \tag{3.5}$$

The angular velocity Ω' of the new cluster is given by

$$\Omega' = \frac{M_c}{J}$$

with

$$M_c = \sum_{i\in C'} m_i X_i \times V_i - \sum_{i\in C'} m_i X_c \times V_c' \quad \text{and} \quad J = \sum_{i\in C'} |X_c - X_i|^2 m_i.$$

The velocity of each sphere of the new cluster after the collision is then given by

$$V_i' = V_c' + \Omega' \times (X_i - X_c),\ i \in C'. \tag{3.6}$$

We consider this dynamics until all particles have come together to form one big cluster.

The collision law described above conserves the linear and the angular momenta. However, in what follows we will use a reduced model which does not incorporate the rotation of the clusters around its center of mass. Consequently, the angular momentum is not conserved. Our aim is to compare different numerical methods for large and dense systems and the rotation movement would bring technical difficulties that do not contribute to the conclusions of this study. The collision rule of the simplified model we consider in this Chapter is given by (3.5) and (3.6) with $\Omega' = 0$.

The algorithms we consider in this study are presented next.

3.2.2 ED SCHEME

An event-driven method integrates exactly equations (3.1). The main difficulty is to compute the exact time at which the next collision event occurs. The idea consists of computing the time of all future collisions between any pair of spheres assuming that these events did not interfere with each other. In other words, one computes the times of collisions between any pair of particles that would collide with each other in a future if there were no other particles in the system. Since the system is bounded, all M clusters can collide with any other cluster, which corresponds to $M(M-1)/2$ collisions. The smallest time is then selected and the system is updated to that time. The formulas used to predict the time of the next collision are deduced in the following [52]. Let X_i' and X_j', be the positions of two particles belonging to two different clusters C_k and C_ℓ at the time $t + \Delta t_{ij}$ of contact. At this time the distance between these two spheres is

$$|X_i' - X_j'| = R_i + R_j. \tag{3.7}$$

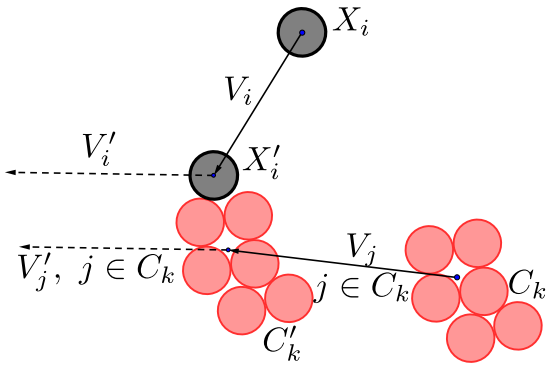


Figure 3.2: Example in two dimensions of a collision between sphere i and cluster k and successive aggregation. The pre and post collisional velocities are represented by solid and dashed lines, respectively.

On the other hand, at time t of the previous event, the positions, X_i and X_j, and velocities, V_i and V_j , of the two spheres satisfy

$$X_i' = X_i + \Delta t_{ij} V_i \text{ and } X_j' = X_j + \Delta t_{ij} V_j. \tag{3.8}$$

Substituting now the $2d$ equations (3.8) into equation (3.7) gives

$$\Delta t_{ij} = \begin{cases} \infty & \text{if} \quad \Delta V \cdot \Delta X \geq 0 \quad \text{or} \quad \sigma < 0 \\ -\frac{\Delta V \cdot \Delta X + \sqrt{\sigma}}{\Delta V \cdot \Delta V} & \text{otherwise,} \end{cases} \tag{3.9}$$

with $\sigma = (\Delta V \cdot \Delta X)^2 - (\Delta V \cdot \Delta V)(\Delta X \cdot \Delta X - (R_i + R_j)^2)$, $\Delta X = X_i - X_j$ and $\Delta v = V_i - V_j$. Using formula (3.9) we compute the time interval Δt_{ij} for each pair (i, j) of spheres that do not belong to the same cluster.Thus, in order to predict the time of the next collision event, it is sufficient to find the minimum between these time increments at time t

$$\Delta t = \min(\Delta t_{ij}),\ \forall\ i, j = 1, ..N,\ i \in C_k,\ j \in C_\ell,\ k \neq \ell. \tag{3.10}$$

In Figure 3.2 it is depicted an example of aggregation between two clusters in two space dimensions together with the pre and post collisional velocities.

Remark 3.2.2. *In order to speed up the algorithm, rather than computing the collision time for all pairs of spheres that do not belong to the same cluster, we only compute it for the pairs of spheres that collide in a near future as follows. We subdivide the domain in boxes of side length ℓ and we compute the future collision times only between the pairs of spheres that lie within the same box or in nearby boxes. Similarly to formula* (3.10)*, we then obtain the smallest time increment $\Delta_1 t$ among all computed increments. In order to avoid missing collisions between spheres that were far from each other we restrict the movement of the spheres between two consecutive iterations of the algorithm to a square of side length $s = \ell/2$. We then compute the maximum time increment $\Delta_s t$ associated to this length as follows: $\Delta_s t = \min_i(s/|V_i|_\infty)$, with $|V_i|_\infty = \max_k |v_{i_k}|$. The time to the next iteration of the algorithm is then given by the minimum between the two increments*

referred above, i.e.,

$$\Delta t = \min(\Delta_s t, \Delta_1 t). \tag{3.11}$$

The system then evolves from event to event as follows. Let the state of the system at time t be given, namely the positions $\mathbf{X}(t) = \{X_i(t)\}_{i=1,...,N}$ and velocities $\mathbf{V}(t) = \{V_i(t)\}_{i=1,...,N}$ of the particles and the sets of particles that belong to the same cluster $C_1(t), ..., C_M(t)$. Additionally, consider the time $t + \Delta t$ of the next event. The state of the system at time $t + \Delta t$ is then obtained by performing the following steps:

1. Evolution of the positions of particles and clusters to time $t + \Delta t$ along a straight line trajectory, i.e.,
$$\mathbf{X}(t + \Delta t) = \mathbf{X}(t) + \Delta t \mathbf{V}(t);$$

2. Actualization of the velocities:

   a) Computation of the new clusters. We determine the number of clusters M and the sets of particles belonging to each cluster $C_1(t + \Delta t), ..., C_M(t + \Delta t)$;

   b) Computation of the velocities of the two colliding clusters according to the collision law (3.5);

3. Computation of the time of the next collision using (3.9) and (3.10).

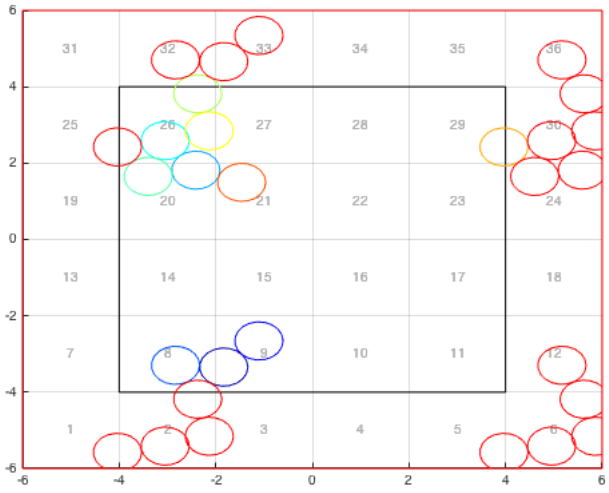


Figure 3.3: Illustration of a 2-dimensional torus divided in boxes with an outer layer of boxes for the ghost spheres (see remark 3.2.3). The real spheres are displayed within the inner square, while the ghost spheres are displayed in red outside the inner square.

Remark 3.2.3. *The spheres are kept inside the domain* $[0, L]^d$ *by applying modulo* L *operation to the positions* $\mathbf{X}$*. The spheres that are near the borders of the hypercube* $[0, L]^d$ *are copied to the opposite border. The copies are called ghost spheres and they are represented in red in the example in Figure 3.3. The ghost spheres are then used to*

compute the Euclidean distances in a torus using formula (3.3). *The collision time* (3.9) *is obtained by considering all (ghost and real) spheres and selecting the smallest time using formulas* (3.10) *and* (3.11).

In the next Section we present the time-discrete dynamics and the TS methods considered in this work.

3.3 Time-discrete model and time-stepping (TS) schemes

3.3.1 Time-discrete model

We begin by reformulating the problem presented in the previous Section as a time-discrete dynamics. We define *apriori* a sequence of time-increments $\Delta t^n > 0,\ n \in \mathbb{N}$ defined iteratively as the largest value such that the movement of all spheres is bounded by a square of side length equal to the radius the sphere R_i, i.e., $\Delta t^n = \min_i R_i/|V_i^{n-1}|_\infty$, with $|V_i|_\infty = \max_k |v_{i_k}|$. This avoids that the spheres pass through each other without any collision being detected. Consider the initial condition $X_i^0 = X_i(0)$ and $V_i^0 = V_i(0),\ i = 1, ..., N$ at time $t^0 = 0$. The sequence of the successive positions $\mathbf{X}^n = \{X_i^n\}_{i=1,...,N},\ n \in \mathbb{N}$ and velocities $\mathbf{V}^n = \{V_i^n\}_{i=1,...,N},\ n \in \mathbb{N}$ of the spheres at time $t^n = t^{n-1} + \Delta t^n$, is obtained by performing a time-discretization of equations (3.1). This operation is then followed by a small correction on the positions or velocities of the spheres to remove the overlapping that might have been introduced in the previous step. This correction typically involves the resolution of a minimization problem. Finally, an actualization of the velocities according to the collision rule (3.5) is performed. An illustration of the time-discrete dynamics is presented in Figure 3.4. In contrast to the time-continuous dynamics, in the time-discrete dynamics multiple collisions may occur simultaneously in the same time-step.

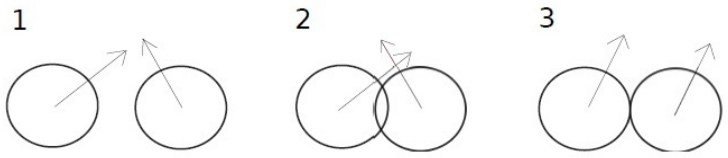


Figure 3.4: Illustration of the three main steps executed during one time-iteration of the TS allgorithms: 1) free motion using a time-discretization of the equations (3.1), 2) adjustment of the positions to correct overlapping, and 3) actualization of the velocities according to collision rule (3.5).

3.3.2 TS schemes

The time-stepping schemes are based on the time-discrete dynamics described above and in the resolution of a minimization problem with non-overlapping constraints at each time-step. In Section 0.3.2.2 we describe some examples of TS schemes that have been proposed in the literature and a general comparison with the methods we propose next.

The state of the system at time t^n is described by the positions $\mathbf{X}^n$ and velocities $\mathbf{V}^n$ of the spheres and the set S^n of pairs of spheres that are linked through adhesive forces. After the minimization algorithm, it may happen that two adhesive spheres end up away from each other, beyond a certain distance, in order to open space for the new spheres arriving in the cluster. In this situation we break the adhesive link between them. The set S^n contains therefore the pairs of spheres with two properties: 1) they came into contact with each other at some time-point prior to t^n, becoming adhesive, and 2) they remained within a certain small distance from each other until time t^n. Let $W_{S^n} : \mathbb{R}^{dN} \to \mathbb{R},\ \mathbf{X} \mapsto W_{S^n}(\mathbf{X})$ be a given real function associated to the set S^n, which will be the function we minimize in the minimization step. Suppose that W_{S^n} has an attainable minimum in the set defined by the non-overlapping constraints defined in (3.2)-(3.4). The general procedure of the time-stepping methods considered in this Chapter is presented next.

Let the initial state of the system $(\mathbf{X}^0, \mathbf{V}^0, S^0)$ at time t^0 be given. Given the state of the system $(\mathbf{X}^{n-1}, \mathbf{V}^{n-1}, S^{n-1})$ at time t^{n-1}, we obtain the state $(\mathbf{X}^n, \mathbf{V}^n, S^n)$ at time $t^n := t^{n-1} + \Delta t^n$, with $\Delta t^n = \min_i R_i/|V_i^{n-1}|_\infty$ by performing the following steps:

1. Free motion. The state of each particle evolves independently of the other particles through an Euler discretization of equation (3.1):

$$\hat{\mathbf{X}}^n = \mathbf{X}^{n-1} + \Delta t^n \mathbf{V}^{n-1};$$

2. Adjustment of the positions to correct overlapping between colliding spheres. We perform the following steps:

   a) Computation of the set of adhesive spheres S^n. Starting from S^{n-1} we add new links corresponding to new contacts and we break links corresponding to spheres whose mutual distance is larger than some threshold:

$$\begin{aligned} S^n = (S^{n-1} \quad & \cap \{(i,j) | i < j,\ \phi_{ij}(\hat{\mathbf{X}}^n) \geq -R_{ij}\epsilon_u\}\ ) \\ & \cup \{(i,j) | i < j,\ \phi_{ij}(\hat{\mathbf{X}}^n) \geq 0\}. \end{aligned} \tag{3.12}$$

   where ϕ_{ij} is the function defined in (3.2), $R_{ij} = \max\{R_i, R_j\}$ and ϵ_u is the threshold for unbinding, which is chosen to be $1/10$.

   b) Finding a set of positions $\mathbf{X}^n$ in the neighbourhood of $\hat{\mathbf{X}}^{n-1}$ that locally solves the minimization problem (see remark 3.3.1):

$$\mathbf{X}^n \in \underset{\phi_{ij}(\mathbf{X}) \leq 0,\ (i,j) \in S}{\operatorname{argmin}} W_{S^n}(\mathbf{X}), \tag{3.13}$$

3. Actualization of the velocities:

   a) Computation of the new clusters. We obtain the number of clusters M^n and the sets of particles belonging to each cluster $C_1^n, ..., C_{M^n}^n$.

b) Computation of the velocities according to collision rule (3.5):

$$\forall k = 1, ..., M^n,\ \forall i \in C_k^n,\ V_i^n = \frac{\sum_{j \in C_k^n} m_j V_j^{n-1}}{\sum_{j \in C_k^n} m_j}.$$

Remark 3.3.1. *During the minimization algorithm in Step 2b), the clusters have to expand in order to eliminate the overlapping in the system. This may lead to new overlapping with nearby clusters that were not overlapping before. Thus, within Step 2, we need to actualize the set of linked (adhesive) spheres and solve the minimization problem again for the new set. We repeat these steps until no more actualizations are needed. The implementation of Step 2 includes therefore an inner and an outer procedure:*

- *The inner procedure stands for the minimization algorithm itself and it will be described in the next two Sections.*
- *In the outer-loop we successively actualize the set of linked spheres $S^{n,q}$ and obtain $\mathbf{X}^{n,q}$ as a solution to the minimization algorithm with the set $S^{n,q}$ until no more actualizations on the set of linked spheres are performed. We start by initializing $\mathbf{X}^{n,0} = \mathbf{X}^n$ and the set $S^{n,0} = S^n$. Let $\mathbf{X}^{n,q-1}$ and $S^{n,q-1}$ be given. At the q^{th} iteration, first we compute $S^{n,q}$. Similarly to* (3.12)*, we define $S^{n,q}$ as the union between the set of the links in S^n that have not been lost during the outer-loop and the set of new links that have been formed, i.e.,*

$$\begin{aligned} S^{n,q} = (S^{n,q-1} \quad & \cap \{(i,j) | i < j,\ \phi_{ij}(\mathbf{X}^{n,q-1}) \geq -R_{ij}\epsilon_u\}) \\ & \cup \{(i,j) | i < j,\ \phi_{ij}(\mathbf{X}^{n,q-1}) \geq 0\}. \end{aligned} \tag{3.14}$$

  Second, if $S^{n,q} = S^{n,q-1}$ then the outer loop stops, otherwise $\mathbf{X}^{n,q}$ is obtained as a solution to the minimization problem for the new set $S^{n,q}$.

The four TS methods considered in this Chapter derive from the next two observations. First, the clusters involved in a collision may either be regarded as a rigid object that moves as a whole, or as a non-rigid object constituted by several spheres that move individually under the effect of adhesion forces and non-overlapping constraints. The two schemes that derive from this observation are described in Sections 3.3.3 and 3.3.4. Second, the non-overlapping constraints may be described by inequalities involving non-smooth or smooth functions, both leading to equivalent constraints. The smooth form of the constraints is given by

$$\phi_{ij}(\mathbf{X}) = (R_i + R_j)^2 - |X_i - X_j|^2. \tag{3.15}$$

and the non-smooth form of the constraints is given by

$$\phi_{ij}(\mathbf{X}) = R_i + R_j - |X_i - X_j|. \tag{3.16}$$

As observed in Chapter 1, the use of each of these functions have differences on the performances of the algorithms. The four different methods result therefore from the combination of using smooth (S) or non-smooth (NS) form of the constraints and rigid (R) or non-rigid (NR) clusters. By combining these two aspects we get the 4 methods we explore in this Chapter that we denote by: S-NR, NS-NR, S-R, NS-R.

The modifications to the original dynamics obtained by substituting the exact collision dynamics with the minimization method give rise naturally to different aggregates of spheres. However, since the original system is highly dependent from the initial position and velocity of the objects which in general cannot be known with precision, we assume that the modification introduced in the dynamics will play the same role of the uncertainty in the system and that this modified system will statistically produce the same type of configurations than the event-driven approach. This assumption should be further documented in future work. The second part of this work is devoted to the numerical comparison between event-driven and time-stepping methods.

3.3.3 Minimization algorithm for non-rigid clusters

In this Section we describe the algorithm that is used in Step 2b) in the previous Section 3.3.2 in the case of non-rigid clusters.

Let $\hat{\mathbf{X}}$ be the configuration with overlapping spheres, called non-admissible configuration, obtained in Step 1 at the n^{th} iteration. Let S be the set of pairs of linked spheres obtained in Step 2a). We want to find a configuration $\bar{\mathbf{X}}$ in the neighbourhood of $\hat{\mathbf{X}}$ such that $\bar{\mathbf{X}}$ does not have overlapping spheres and the pairs of linked spheres remain near each other. This can be formulated as the search for a configuration $\bar{\mathbf{X}}$ in the neighbourhood of $\hat{\mathbf{X}}$ that locally minimizes the sum of the squared distances between the pairs in S i.e.,

$$W_S(\mathbf{X}) = \frac{1}{2} \sum_{(i,j)\in S} |X_i - X_j|^2,$$

under non-overlapping constraints $\phi_{ij}(\mathbf{X}) \leq 0,\ (i,j) \in S$. Note that in contrast to the minimization problem considered in Part I, the function that we minimize represents a local potential, i.e., each particle interacts only with their neighbouring particles. Similarly to the problems studied in Part I, this problem has also multiple solutions and a local solution can be computed by the 1st-order DAHA that was developed in Chapter 1. We briefly describe it next.

Let $\#S$ be the number of pairs of linked spheres. Consider the Lagrangian $\mathcal{L}_S : \mathbb{R}^{dN} \times (\mathbb{R}_0^+)^{\#S}$ defined by

$$\mathcal{L}_S(\mathbf{X}, \boldsymbol{\lambda}) = W_S(\mathbf{X}) + \sum_{(i,j)\in S} \lambda_{ij}\phi_{ij}(\mathbf{X}) \tag{3.17}$$

where $\boldsymbol{\lambda} = \{\lambda_{ij}\}_{(i,j)\in S}$ is the set of Lagrange multipliers associated to the non-overlapping constraints. Consider the first-order damped Arrow-Hurwicz system (DAHS) [45]

$$\begin{cases} \dot{X}_k & = -\alpha^2 \nabla_{X_k} \mathcal{L}_S(\mathbf{X}, \boldsymbol{\lambda}) - \gamma^2 (\sum_{(i,j)\in S} \phi_{ij}(\mathbf{X}) \lambda_{ij} \nabla_{X_k} \phi_{ij}(\mathbf{X})) \\ & \qquad\qquad\qquad\qquad\qquad\qquad k = 1, ..., N \qquad (3.18) \\ \dot{\lambda}_{ij} & = \beta \nabla_{\lambda_{ij}} \mathcal{L}_S(\mathbf{X}, \boldsymbol{\lambda}), \quad (i,j) \in S \qquad (3.19) \end{cases}$$

where the dot represents the derivative with respect to a fictitious time and α, β and γ are positive parameters.

Substituting (3.17) into (3.18)-(3.19), we obtain

$$\begin{cases} \dot{X}_k & = -\alpha^2 (\nabla_{X_k} W_S(\mathbf{X}) + \sum\limits_{(i,j)\in S} \lambda_{ij} \nabla_{X_k} \phi_{ij}(\mathbf{X})) \\ & \qquad -\gamma^2 (\sum_{(i,j)\in S} \phi_{ij}(\mathbf{X}) \lambda_{ij} \nabla_{X_k} \phi_{ij}(\mathbf{X})) \\ & \qquad\qquad\qquad\qquad k = 1, ..., N \qquad (3.20) \\ \dot{\lambda}_{ij} & = \begin{cases} 0, & \text{if } \lambda_{ij} = 0 \text{ and } \phi_{ij}(\mathbf{X}) < 0 \\ \beta \phi_{ij}(\mathbf{X}), & \text{otherwise} \end{cases}, \quad (i,j) \in S. \qquad (3.21) \end{cases}$$

From Theorem 4 presented in Section 0.3.2.1, the set of solutions to the minimization problem is contained into the set of critical-points of the Lagrangian. These critical points coincide with the steady states of the DAHS.

Indeed, from equation (3.21) we have that a steady state $(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}})$ satisfies $\bar{\lambda}_{ij} \phi_{ij}(\bar{\mathbf{X}}) = 0$. Consequently, from the system (3.18)-(3.19) a steady state satisfies $\nabla_{X_k} \mathcal{L}_S(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}}) = 0$, and $\nabla_{\lambda_{ij}} \mathcal{L}_S(\bar{\mathbf{X}}, \bar{\boldsymbol{\lambda}}) = 0$, $(i,j) \in S$, $k = 1, ..., N$. The 1st-order DAHA consists on a semi-implicit discretization of the DAHS system:

$$\begin{cases} X_k^p & = X_k^{p-1} - \alpha^2 (\nabla_{X_k} W_S(\mathbf{X}^{p-1}) + \sum\limits_{(i,j)\in S} \lambda_{ij}^{p-1} \nabla_{X_k} \phi_{ij}(\mathbf{X}^{p-1})) \\ & \qquad -\gamma^2 (\sum\limits_{(i,j)\in S} \phi_{ij}(\mathbf{X}^{p-1}) \lambda_{ij}^{p-1} \nabla_{X_k} \phi_{ij}(\mathbf{X}^{p-1})) \\ & \qquad\qquad\qquad\qquad k = 1, ..., N \qquad (3.22) \\ \lambda_{ij}^p & = \max\{0, \lambda_{ij}^{p-1} + \beta \phi_{ij}(\mathbf{X}^p)\}, \quad (i,j) \in S, \qquad (3.23) \end{cases}$$

where α, β and γ are positive numerical parameters. The convergence test of this scheme reads

$$\frac{|\mathbf{X}^p - \mathbf{X}^{p-1}|}{|\mathbf{X}^{p-1}|} \leq \epsilon_X, \qquad (3.24)$$

for a small positive constant ϵ_X.

As a final remark note that a steady state of the DAHS (3.20)-(3.21) satisfies

$$\nabla_{X_k} W(\bar{\mathbf{X}}) = -\sum_{(i,j)\in S} \lambda_{ij} \nabla_{X_k} \phi_{ij}(\bar{\mathbf{X}}), \ k = 1, ..., N. \qquad (3.25)$$

These equalities show that a solution to the minimization problem corresponds to a balance between attraction forces associated to the adhesive particles (left hand-side) and repulsion forces associated to overlapping particles (right hand-side). The dynamics of the DAHA algorithm corresponds thus to a search for this balance.

3.3.4 Minimization algorithm for rigid clusters

In the previous approach, the relative position of the spheres in one cluster may change over time after each minimization problem. In this Section we propose an algorithm that ensures that the non-overlapping linked spheres keep the same relative position. In this way, the clusters are kept rigid and the dynamics of the system should get closer to the dynamics produced by the ED described in Section 3.2.2.

We start by reformulating the minimization problem appearing in Step 2b) in Section 3.3.2. Let $\hat{\mathbf{X}}$ be the configuration with overlapping spheres obtained in Step 1 at time t^n. Let $C_1^0, ..., C_M^0$ be the sets of spheres belonging to the same cluster at the end of the previous time-point t^{n-1}. Consider the set of overlapping spheres S_{overl} defined by

$$S_{\text{overl}} = \{(i,j) \mid \phi_{ij}(\hat{\mathbf{X}}) \geq 0\}. \tag{3.26}$$

and the set of rigid configurations with respect to $\hat{\mathbf{X}}$ defined by

$$\mathcal{X}(\hat{\mathbf{X}}) := \{\mathbf{X} \in \mathbb{R}^{dN} | \ \forall k \in 1, ..., M, \ \forall i,j \in C_k^0, \ X_i - X_j = \hat{X}_i - \hat{X}_j\}.$$

This set contains all configurations such that the spheres belonging to the same cluster at time t^{n-1} keep the same relative positions they had at time t^{n-1}. The minimization problem is similar to the one formulated in the previous approach, except that instead of using the pairs of linked spheres we use the pairs of overlapping spheres S_{overl} and we look for solution within $\mathcal{X}(\hat{\mathbf{X}})$. In other words, we search for a configuration $\bar{\mathbf{X}} \in \mathcal{X}(\hat{\mathbf{X}})$ in the neighbourhood of $\hat{\mathbf{X}}$ that locally minimizes the function

$$W_{S_{\text{overl}}}(\mathbf{X}) = \frac{1}{2} \sum_{(i,j) \in S_{\text{overl}}} |X_i - X_j|^2 \tag{3.27}$$

subject to non-overlapping constraints

$$\phi_{ij}(\mathbf{X}) \leq 0, \ i,j \in S_{\text{overl}}.$$

To solve this problem, we consider the Lagrangian $\mathcal{L}_{S_{\text{overl}}}$ as defined in (3.17) and a modification of the DAHS (3.20)-(3.21) in which all spheres in the same cluster are subjected to the same dynamics. Specifically, they are subjected to the sum of the forces

acting in each particle of the cluster. The resulting dynamical system is then given by

$$\begin{cases} \dot{X}_p &= -\alpha^2 \sum_{\ell \in C_k} \nabla_{X_\ell} \mathcal{L}_{\text{overl}}(\mathbf{X}, \boldsymbol{\lambda}) \\ & \qquad -\gamma^2 \sum_{\ell \in C_k} \Big(\sum_{(i,j) \in S_{\text{overl}}} \phi_{ij}(\mathbf{X}) \lambda_{ij} \nabla_{X_\ell} \phi_{ij}(\mathbf{X}) \Big) \\ & \qquad\qquad\qquad\qquad \forall p \ \in C_k, \ k = 1, ..., M \qquad (3.28) \\ \dot{\lambda}_{ij} &= \begin{cases} 0, & \text{if } \lambda_{ij} = 0 \text{ and } \phi_{ij}(\mathbf{X}) < 0 \\ \beta \phi_{ij}(\mathbf{X}), & \text{otherwise} \end{cases} , \ (i,j) \in S_{\text{overl}}. \qquad (3.29) \end{cases}$$

Similarly to the case of non-rigid clusters, we then consider a semi-implicit discretization of this system and the convergence criterion (3.24).

3.4 Numerical study

In this Section we compare the ED and TS algorithms described in the previous Sections. We first tune the parameters of the minimization algorithms in order to maximize the efficiency of the TS methods. Then we compare the TS and ED methods in terms of computational time for different values of N up to $N = 10^6$ and for two different values of the size of the domain.

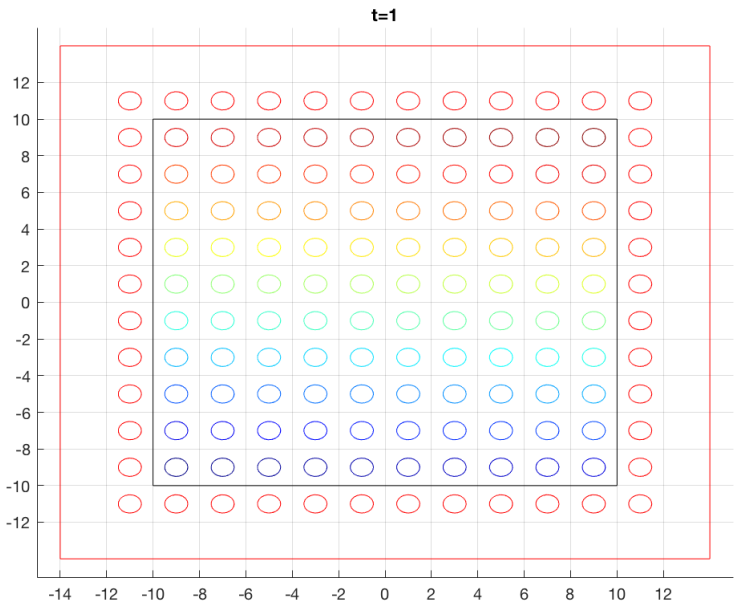


Figure 3.5: Initial condition with $N = 100$ particles.

The numerical simulations are performed in dimension $d = 2$. The experimental set up used in this study is the following. We consider N spheres with the same mass $m_i = m = 1$ and radius $R_i = R$, $i = 1, ..., N$, given by $R = 3/(4\sqrt{N})$. The spheres move in a 2-dimensional torus $[0, L]^2$ divided in N squared boxes with side length $\ell = \ell_1 R$.

In the following we use $\ell_1 \in \{2.5, 3\}$ and $N \in [10^2, 10^6]$. Note that the volume fraction occupied by the spheres in the situation described above is relatively high, namely, 0.5 for $\ell_1 = 2.5$ and 0.35 for $\ell_1 = 3$. The initial condition is given by a configuration where each sphere is located at the center of each box as in Figure (3.5) and the initial velocities are randomly generated from the uniform distribution with support $[0, V_{\max}]$, with $V_{\max} = 5R$.

In the next Sections, for each of the quantities plotted, we represent the average value over the different initial conditions and we use error bars to represent the standard deviation[‡]. The ED method is represented by a black line. The TS methods with smooth or non-smooth constraints are represented by a green or red line, respectively. The TS methods with rigid or non-rigid clusters are represented by a dashed or solid line, respectively.

3.4.1 Choice of parameters of minimization algorithms

In this Section we consider the four minimization algorithms: S-NR, NS-NR, S-R, S-NR and we choose the most appropriate parameters α, β, γ and ϵ_X which are going to be used in the comparative study presented in the next Section. We first consider $\epsilon_X = 10^{-6}$ and explore the parameter space for α, β and γ. We choose the values that lead to a small number of iterations of the minimization algorithm. Then we use these values to run a similar study for the tolerance ϵ_X. We choose the value for ϵ_X that leads to a small number of iterations of the minimization algorithm while preserving the accuracy of the system. Notice that the choice of the parameters is going to depend on the minimization method and on the parameters of the model, such as the number of spheres N and the radius R.

3.4.1.1 Choice of α, β and γ

We first run simulations to choose the parameters α, β and γ for each of the four minimization algorithms. The parameter α is related to the speed of actualization of the positions of the spheres in the direction of minus the gradient of the potential W_S while the parameter β controls how fast the Lagrange multipliers associated to the constraints are actualized (see Part I). Consequently, α is related to the intensity of the attraction forces and β to the intensity of the repulsion forces. The larger the values of these parameters, the faster the dynamics of the system, but the harder it is to converge (see Chapter 2). So we have to find intermediate values which lead to a fast convergence. For each value of the parameters, we compute the average number of iterations of the minimization algorithm $N_{\text{iter}}^{\text{mean}}$ over the number of simulations during the time-evolution of the system,

[‡]The simulations were run on a Dell PowerEdge R630 rackmount server fitted with two Intel Xeon E5-2637 3.5 GHz processors, each of which with 4 cores and 8 threads. The results were plotted using matlabR2014b.

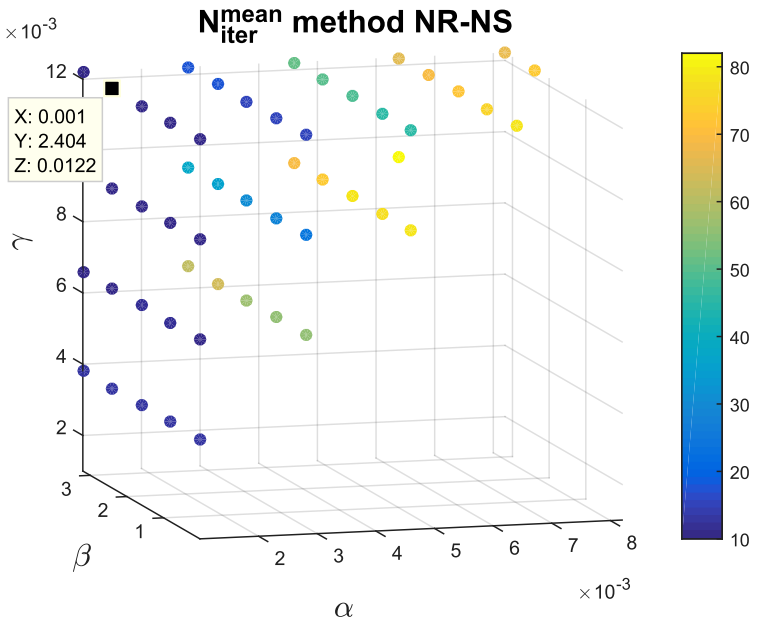


Figure 3.6: Plot of $N_{\text{iter}}^{\text{mean}}$ (in color code) as functions of the parameters α, β, γ for the NS-NR algorithm with $N = 10^4$, $R = 3/(4\sqrt{N})$ and $\ell = 2.5R$. Dark blue corresponds to $N_{\text{iter}}^{\text{mean}} = 10$ and yellow corresponds to $N_{\text{iter}}^{\text{mean}} = 80$. The parameters $(\alpha, \beta, \gamma) = (0.001, 2, 0.01)$ were chosen from the blue region.

i.e.,

$$N_{\text{iter}}^{\text{mean}} = \frac{1}{N_{\text{simu}}} \sum_{n=1}^{N_{\text{simu}}} N_{\text{iter}}^{n},$$

where N_{simu} is the total number of simulations and N_{iter}^{n} is the number of iterations of the minimization algorithm at the n^{th} simulation.

As an example, in Figure 3.6 we show the results in the case of NS-NR with $N = 10^4$ and $\ell = 2.5R$. The graph shows $N_{\text{iter}}^{\text{mean}}$ (in color code) as a function of α, β, γ, where $\alpha \in [0.001, 0.008]$, $\beta \in [0.01, 3.2]$ and $\gamma \in [0.004, 0.012]$. Notice that in order to simplify the visualization we only plot the dots corresponding to $N_{\text{iter}}^{\text{mean}} < 82$. The color code ranges from dark blue, with $N_{\text{iter}}^{\text{mean}} = 10$, to yellow, with $N_{\text{iter}}^{\text{mean}} = 82$. We then choose the values of the parameters within the blue region. In this case we have chosen $(\alpha, \beta, \gamma) = (0.001, 2, 0.01)$.

This study has been conducted for each algorithm and several values of N. From these simulations we have been able to extract the parameters which are presented in Table 3.1. The algorithms with rigid cluster use the same parameters as the algorithms with non-rigid clusters. In the case of the smooth methods, the parameters have spatial units, therefore, they depend on R.

3.4.1.2 Choice of ϵ_X

Once the parameters α, β and γ have been chosen, we determine the value of the tolerance. The idea is to choose a tolerance ϵ_X such that the number of iterations of the minimization is low while the error, which measures how well the hard constraints are verified, remains

N	α	β	γ
smooth constraints			
-	0.001	$0.01/R^2$	$0.1/R$
non-smooth constraints			
9×10^2	0.003	2	0.1
$[2.5, 6.4] \times 10^3$	0.002	2	0.02
$[1, 4] \times 10^4$	0.001	2	0.01
$[1.6, 2.5] \times 10^5$	0.001	2	0.003
10^6	0.0001	2	0.001

Table 3.1: Parameters for the minimization problems with smooth and non-smooth constraints for different values of N and for $R = \frac{3}{4\sqrt{N}}$ and $\ell = 2.5R$.

small. This parameter is going to be fixed for all of the following simulations and it is not going to depend on the algorithm used nor on the number of particles. To compare the influence of the tolerance, we are going to compute the average number of iterations, $N_{\text{iter}}^{\text{mean}}$, and the error, $E^{textmean}$, which is given by

$$E^{\text{mean}} = \frac{1}{N_{\text{simu}}} \sum_{n=1}^{N_{\text{simu}}} \frac{1}{\#S^n} \sum_{(i,j)\in S^n} \frac{|\phi_{i,j}(X)|}{\mathcal{R}},$$

where S^n contains the pairs of linked spheres at the n^{th} minimization as defined in (3.12) and $\#S^n$ is the number of elements in S^n. The function ϕ_{ij} can have the smooth or non-smooth form as defined in (3.15) and (3.16), respectively, and the normalization constant $\mathcal{R}$ takes either the value $2R$ or $4R^2$ depending on the function ϕ_{ij} used. Figure 3.7 shows E^{mean} and $N_{\text{iter}}^{\text{mean}}$ as a function of the tolerance $\epsilon_X = 10^{-n}$ with

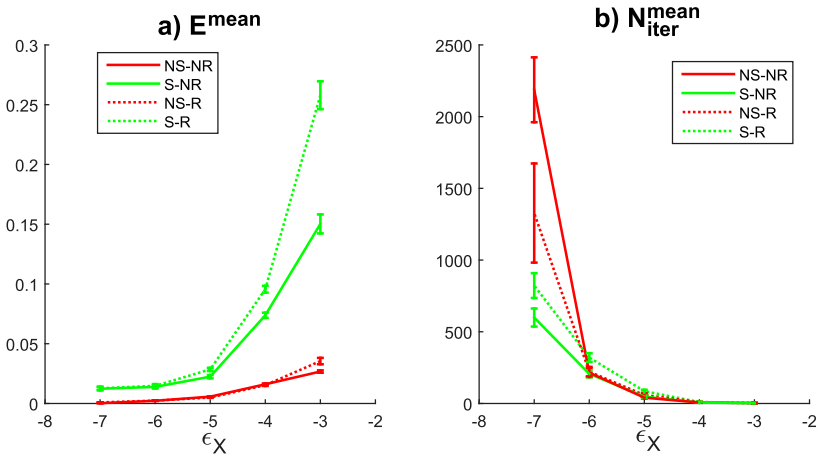


Figure 3.7: Plot of E^{mean} (a) and $N_{\text{iter}}^{\text{mean}}$ (b) as functions of the tolerance $\epsilon_X \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ for NS-NR (red solid line), S-NR (green solid line), NS-R (red dashed line), S-R (green dashed line) with $N = 900$, $R = 3/(4\sqrt{N})$, $\ell = 2.5R$ and for 10 initial conditions. The value of the tolerance that minimizes the number of iterations and the error is $\epsilon = 10^{-6}$.

$n \in \{-7, -6, -4, -3, -2, -1\}$. The simulation has been run for the four methods S-NR, NS-NR, S-R, NS-NR and 10 initial conditions with $N = 900$, $\ell = 2.5R$ and the

remaining parameters taken from table 3.1.

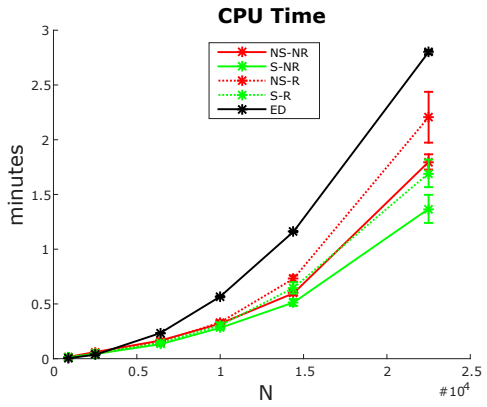
In Figure 3.7a) we observe that the error increases with the tolerance, which is expected. We consider that the hard constraint is verified if the error is around 10^{-3}, which occurs for a tolerance equal to 10^{-7} or 10^{-6}. However in Figure 3.7b) we observe that the mean number of iterations of the minimization decreases exponentially with the tolerance. From $\epsilon_X = 10^{-7}$ to $\epsilon_X = 10^{-6}$ the mean number of iterations ranges from two times smaller in the case of the S-NR and one hundred times smaller in the case of the NS-NR. Since we want to keep the number of iteration as low as possible we choose the tolerance $\epsilon_X = 10^{-6}$.

3.4.2 Comparison between TS and ED schemes

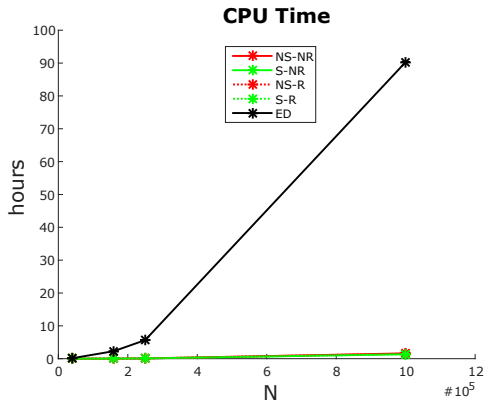
In this Section we compare the ED and TS schemes regarding the computational time they take to simulate the phenomenon of ballistic aggregation described in Section 3.2. In particular, we study the influence of the number of spheres N and of the size of the domain. We use the numerical parameters obtained in the previous Section, namely, α, β and γ presented in Table 3.1 and the tolerance $\epsilon_X = 10^{-6}$.

3.4.2.1 Influence of the number of particles

We first compare the TS algorithms with the ED algorithm with respect to the computational time for several values of N. In this Section we consider a set of 5 initial conditions and like in the previous Section, we use the radius $R = 3/(4\sqrt{N})$ and the side of a box $\ell = 2.5R$. For these values, the volume occupied by the spheres and the volume of the domain remain constant over N. As we will see, the computational time of the ED becomes larger than the TS schemes as N increases.



(a) $N \in \{0.9, 2.5, 6.4, 10, 14.4, 22.5\} \times 10^3$.



(b) $N \in \{4, 1.6, 2.5, 10\} \times 10^5$.

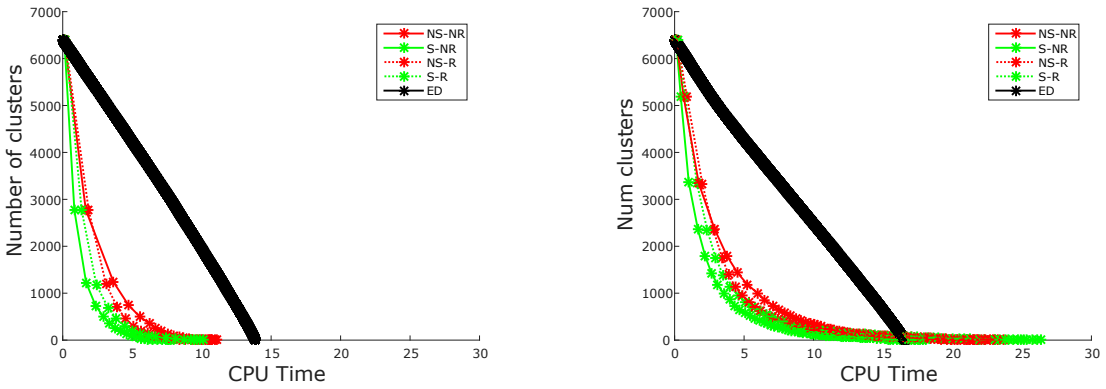
Figure 3.8: Computational time as a function of N for NS-NR (red solid line), S-NR (green solid line), NS-R (red dashed line), S-R (green dashed line), ED (black line) and 5 initial conditions with $R = 3/(4\sqrt{N})$ and $\ell = 2.5R$. We observe that the TS algorithms outperform the ED for $N \geq 6400$.

We obtain the computational time (CPU time), for each method and for several values of N, from $N = 900$ up to $N = 10^6$. In Figure 3.8 we present the CPU time as a function of

N for four time-stepping algorithms (S-NR, NS-NR, S-R, NS-R) and for the event-driven. We observe that the ED needs less time than the TS for $N \leq 900$ and the TS algorithms overcome the ED for $N \geq 6400$. In particular, for $N = 10^6$ the ED takes almost four days, while the TS algorithms take less than two hours. When we compare the TS algorithms between themselves we do not see such a big difference. The algorithms with smooth form of the constraints (in green) seem to perform better than with the non-smooth form (in red). The algorithms with non-rigid cluster (in solid line) seem to perform better than with rigid cluster (dashed line). A possible explanation for this last observation could be that a rigid cluster may have more difficulty in finding space in a crowded environment than a non-rigid one.

3.4.2.2 Influence of the size of the domain

We now fix the number of spheres $N = 6400$ and analyse the effect of increasing the size of the domain. Specifically, we consider the side length of the box that we have been using, namely, $\ell = 2.5R$ and a larger one $\ell = 3R$. As we will see, the TS algorithms are negatively affected by this increase on the size of the domain, while the ED algorithm is not substantially affected by this change.



(a) Side length of a box $\ell = 2.5R$.

(b) Side length of a box $\ell = 3R$.

Figure 3.9: Number of clusters over computational time during one simulation for NS-NR (red solid line), S-NR (green solid line), NS-R (red dashed line), S-R (green dashed line), ED (black line), for two different values of length side of a box $\ell = 2.5R$ (a) and $\ell = 3R$ (b) with $N = 6400$ and $R = 3/(4\sqrt{N})$. We observe that the ED takes longer in the first case and less time in the second case compared to the TS algorithms.

For each value of ℓ, we plot in Figure 3.9 the number of clusters that are being formed after each time-iteration as a function of the computational time for one initial condition. Naturally, we observe in all schemes that the number of clusters is monotonically decreasing from N, at the beginning of the simulation, to 1, at the end of the simulation. In the ED scheme we observe that the number of clusters decreases linearly over computational time. This means that the computational time that it takes to reduce the number of clusters by one (i.e., to compute one collision) is independent on the number of clusters existing in the system. Each star in the graph represents one time-iteration. Looking at

the first star of the TS schemes in Figure 3.9a), we observe that the number of clusters has dropped more than 3000, i.e., more than 3000 collisions have occurred during the first time-step. After that, the number of collisions per time-step reduces drastically. This may be explained by the decrease on the number of clusters over time. Indeed, less clusters implies a lower rate of collisions. Consequently, less collisions are solved at one time-step and the TS algorithms become less efficient. At some point during the simulation, the TS schemes become less efficient than the ED.

If we consider a larger domain then the overall rate of collisions decreases and the TS schemes may lose against the ED, as we see in Figure 3.9b) where the side length of a box is $\ell = 3R$. This shows that, for a fixed N, there is a threshold on ℓ (and consequently on the size of the domain), where the TS overcomes the ED. This threshold lies between $2.5R$ and $3R$. In particular, we deduce that the fastest algorithm would result from a combination between the two types of schemes: a TS scheme to simulate the first part of the dynamics where the rate of collisions is high and an ED scheme to simulate the second part of the dynamics where the rate of events is below some threshold. In order to develop such a hybrid scheme, the best switching point would have first to be found. This investigation is left for future work.

3.5 Conclusion and future work

In this work we propose four different time-stepping schemes to simulate ballistic aggregation in a d-dimensional torus $[0, L]^d$ in the case where the volume fraction occupied by the particles is high. The TS schemes involve the resolution of a minimization problem with non-overlapping constraints at each time-step which is solved by the damped Arrow-Hurwicz algorithm developed in Part I. The TS algorithms differ from each other on the way the growing aggregates are modelled (rigid or non-rigid) and on the form of the non-overlapping constraints (smooth or non-smooth). We compare the four methods with an event-driven algorithm in terms of computational time. In particular, we study numerically the influence of the number of spheres N and of the size of the domain in dimension $d = 2$ for equally sized spheres with radius R. We observe that the TS methods get increasingly faster than the ED as N increases. For a size of the domain of $L = \ell\sqrt{N}$, with $\ell = 2.5R$, the TS overtakes the ED for $N \geq 6400$. In particular, for $N = 10^6$ particles the ED takes almost four days, while the TS takes only a few hours. On the other hand, the TS methods become slower as ℓ increases. Specifically, for $N = 6400$ particles, the threshold on ℓ at which the TS method becomes slower than the ED lies within $2.5R$ and $3R$. By plotting the evolution of the number of clusters over computational time during one simulation, we observe that the TS scheme is very efficient at the beginning of the dynamics, as many collisions are solved simultaneously at one time-step. In contrast, at the end of the simulation the TS schemes become very inefficient. This indicates that the fastest algorithm to simulate the dynamics should combine a TS at the first part of

the dynamics with an ED at the second. This observation should be further investigated in the future, namely, at which point of the dynamics should one switch between the TS and the ED. Another important aspect that should be explored in the future is how well the TS schemes describe the real dynamics of the system. This may be accomplished by assuming the ED as a reference point for accuracy and comparing the dynamics obtained by the TS with the ED. After selecting an efficient and accurate scheme, one could then study statistical properties associated to the growing aggregates, namely, how many aggregates of size x there are at time t or which type of shapes of the final aggregate may emerge for each type of initial conditions one may consider.

4

Modelling a packed cell tissue with noise

In a packed tissue, neighbouring cells exert high pressure on each other at all times. Such mechanical interactions are believed to play an important role on the dynamics and shape of the tissue. However, their exact contributions are still not yet fully understood. To explore this relation we develop a 2D individual based model for the particular case of an epithelial tissue where the nucleus is the most rigid part of the cell, leading to nuclei packing. The model is based on a geometric representation of individual cells through soft-hard-spheres, modelling the nuclei, and springs, modelling the cytoplasm. The cells interact with each other aiming at minimizing a local potential energy at all times, subject to non-overlapping constraints. This problem is formulated as a non-convex minimization problem, which is tackled with the damped Arrow-Hurwicz algorithm developed in Part I. The evolution of the system is then triggered by cell division, noise and changes in cell characteristics and it is simulated by the time-stepping scheme described in Chapter 3 with smooth constraints. Numerical results show good agreement with experimental data, which has allowed us to support some hypotheses previously proposed by biologists regarding the underlying mechanics of the tissue. In particular, the model suggests that the presence of noise in the system is essential to allow the movement of nuclei in a crowded environment. Preliminary results suggest that the height of the tissue may be directly related to the speed at which the nuclei move inside the cells.

This work is based on a collaboration between mathematicians Pierre Degond[*] and Sara Merino-Aceituno[†] and a team of biologists[‡] led by Eric Theveneau and including Fernando Duarte.

[*]Department of Mathematics, Imperial College London, UK
[†]School of Mathematical and Physical Sciences, University of Sussex, UK
[‡]Centre de Biologie du Développement (CBD), University Paul Sabatier, CNRS, France

4.1 Introduction

Mathematical modelling has been used as a low-cost research tool in the study of the mechanics of cell tissues, as a complement to laboratory experiments. In particular, some questions have recently been explored by resorting to this tool, such as, how a change in spatial constraints affects the cell cycle [170], how curvature of an epithelial sheet is determined by mechanical tensions [84] or what are the sources of the forces acting in each cell within the tissue [182]. We refer to Section 0.2.2.3 for a brief overview of tissue mechanics models. The general modelling methodology consists of building a model based on hypotheses proposed by biologists and observing the model predictions. If these predictions are in agreement with experimental data, then they might be kept for further exploration. If not, then they might be incomplete or wrong and new hypotheses should be formulated. In this work, we focus on packed cell tissues. The dynamics of such tissues is strongly influenced by the contact forces between neighbouring cells. Our aim is to develop a modelling framework which can be used to predict the impact of these forces on the dynamics and shape of the tissue. Such a framework is developed in the context of the neroepithelium of a chick embryo, which is described in the next Section.

The outline of the Chapter is as follows. In Section 4.2 we provide the biological background and the questions that motivate this work. In Section 4.2.2 we justify our modelling choices. In Section 4.3 we describe the mathematical model of the tissue at equilibrium as a solution to a minimization problem. In Section 4.4 we introduce perturbations to this equilibrium state, such as cell division, noise and changes in cells characteristics, and obtain a model for the time-evolution of the tissue. In Section 4.5 we present the numerical results as follows. We start in Section 4.5.1 by estimating the ranges for the parameters of the model from experimental data. Then in Section 4.5.2 we present the influence of the magnitude of the forces and of the intensity of noise. Finally, in Section 4.5.4, statistical results for the evolution of the shape and morphology of the system over 36 hours are presented and compared with experimental data. Finally, the model results are discussed in Section 4.6 and possible future directions and applications are described in Section 4.7.

4.2 Biological background and modelling principles

4.2.1 The neuroepithelium

In this Section we present the main features of the chick neuroepithelium which is used as a biological model in this study.

The *epithelial tissue* or *epithelium* is one of the basic types of animal tissue and it is

constituted of closely packed cells adjacent to a basement membrane. The epithelium can be classified according to the number of cell layers, into simple, pseudostratified (one layer) or stratified (several layers), and according to the shape of the cells, into squamous, cuboidal or columnar.

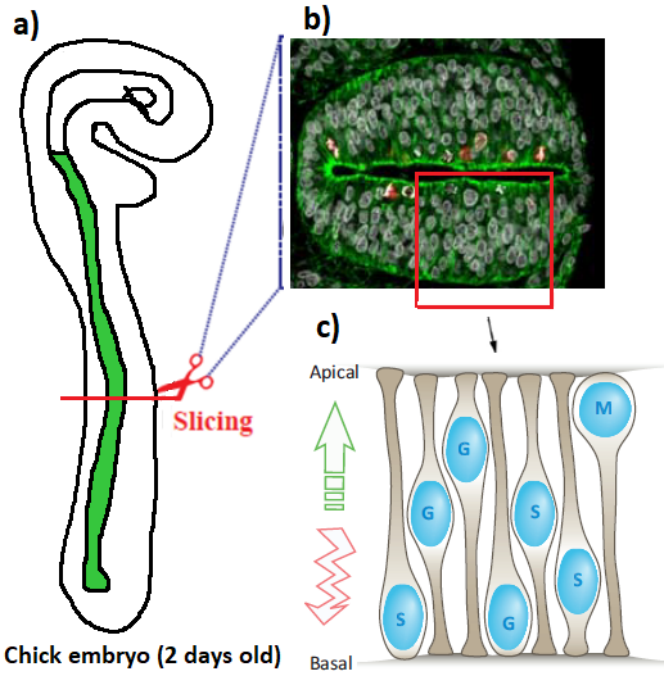


Figure 4.1: Figure a) represents a 2-days-old embryo with the neural tube in green. Figure b) was obtained by Theveneau's lab and it contains a cross section of the neural tube represented in Figure a). The apical network is located in the inner part of the tube, while the basement membrane in the outer part. In between the two we see the nuclei in white. Figure c) was taken from [109] and it contains a sketch of a subsection of Figure b) representing the morphology of a pseudostratified epithelium. Each cell has an elongated shape and it is attached at one side to the basement membrane and at the other side to the apical network. The nucleus is the most rigid part of the cell. Prior to division the nucleus migrates to the apical side as represented in the right most cell in Figure c). The tissue grows in all three directions due to cell division and cell stretching.

The tissue we are interested in is the neuroepithelium which constitutes the neural tube of the chick embryo (see Figure 4.1(a) and (b)). This tissue belongs to the class pseudostratified columnar epithelial tissue, as it contains only one layer of elongated column-shaped cells whose nuclei can be found at different heights inside the cell, giving the perception of a multilayer tissue. In the neuroepithelium of the chick embryo, we observe up to 8 layers in a two-days old embryo, as it is shown in Figure 4.2. The cells are adjacent to a basement membrane. Epithelial cells have polarity, at the basal side the cells are attached to the membrane and at the apical side they are attached to each other, forming an apical network (see Figure 4.1(c)). The volume of a cell is nearly the volume of the nucleus. The nucleus is also the most rigid part of the cell.

The nucleus moves inside the cell along the apico-basal axis. This is called *interkinetic nuclear migration, INM*. The cells are continuously getting stretched, leading to the growth of the tissue in the apico-basal direction. At some point of the cell cycle,

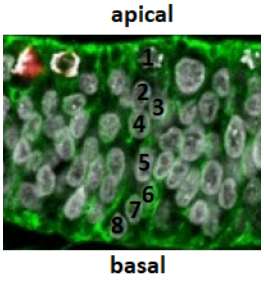


Figure 4.2: A $2D$ image of a section of the neuroepithelium of a two-days old chick embryo. We can see that there are approximately 6 to 8 layers of nuclei. The green line on top corresponds to the apical network and the green line in the bottom corresponds to the basement membrane. The red nucleus near the apical network is in the process of division. Image obtained by Theveneau's lab.

the nucleus tries to approach the apical side. At this stage, the nucleus movement is considerably faster than before. After a while, the nucleus gets more rigid and then it divides. Nearly all divisions are observed to occur when the nucleus is at the apical side of the tissue. All nuclei are constantly shaking without having any apparent directed movement, which may be a consequence of internal sub-nuclear processes. This shaking behaviour may be necessary to allow the rapid movement of nuclei along the apico-basal direction prior to division. Due to crowding, one daughter cell may be born at a different cross-section than the mother cell, leading to the growth of the tissue in the longitudinal direction of the tube.

The mechanical interactions occurring within the neuroepithelium are believed to play a crucial role on the dynamics and architecture of this tissue [109]. Specifically, a few questions one may ask are:

- How does a macroscopic structure, namely the neural tube, emerge and how does it remain stable, despite all the complex processes occurring at the cell level, such as INM and crowding.
- How does the nucleus migrate to the apical side prior to division, despite the crowding and the movement of neighbouring nuclei.
- What determines the number of layers of nuclei along the apico-basal axis and how does that affect INM.

The questions referred above are especially relevant in two contexts: embryonic development and cancer formation. In the context of embryonic development, one tries to understand how organs acquire their shape and how it remains stable [83]. In the context of cancer, how the shape is disrupted and how this disruption allows cancer cells to leave the tissue and start spreading to other organs [172, 98].

The mechanics underlying cell tissues have been addressed in the literature by complementing lab experiments with mathematical modelling and computer simulations, as some hypothesis are very difficult to test in the lab. This is what we are going to do as

well. In this first stage of the project, we have developed a 2D model for the neuroepithelium based on hypotheses formulated from the observed cell behaviour and we have validated the model against experimental data. In a second stage we will perform *in silico* experiments to try to get a deeper insight about the interplay between crowding, INM, tissue shape and growth.

4.2.2 Modelling principles

In this Section we present and justify our modelling approach.

We want to study the relationship between intercellular forces and the dynamics and shape of the tissue, therefore a multiscale approach would be appropriate. The general reasoning that motivates our modelling choices is described in Figure 4.3. The main aim is to be able to derive two models of the tissue at different spatial scales, called the *micro-* and the *macroscopic* models, that are mathematically linked, in the sense that the macroscopic model can rigorously be derived from the microscopic one. Such a mathematical framework, would give a powerful tool in the study of biological systems, as it would allow to enlarge and reduce the spatial scale as needed, just like a "mathematical microscope". This feature also allows to reduce the computational cost. Moreover, a rigorous bridge between two spatial scales would allow to keep consistency between variables and constants across scales, which is very relevant in biology, since the data is usually obtained from macroscopic scales, but the mechanisms that explain these data often occur at microscopic scales. In sum, in order to be able to fully understand a biological system, a rigorous multiscale approach is often needed. Due to the difficulty of the problem, it is desirable to consider the simplest possible individual based model (IBM), which we describe next. A macroscopic model corresponding to a simplified IBM is developed in Chapter 5. Nevertheless, the link between the two models is still missing and is left for future work.

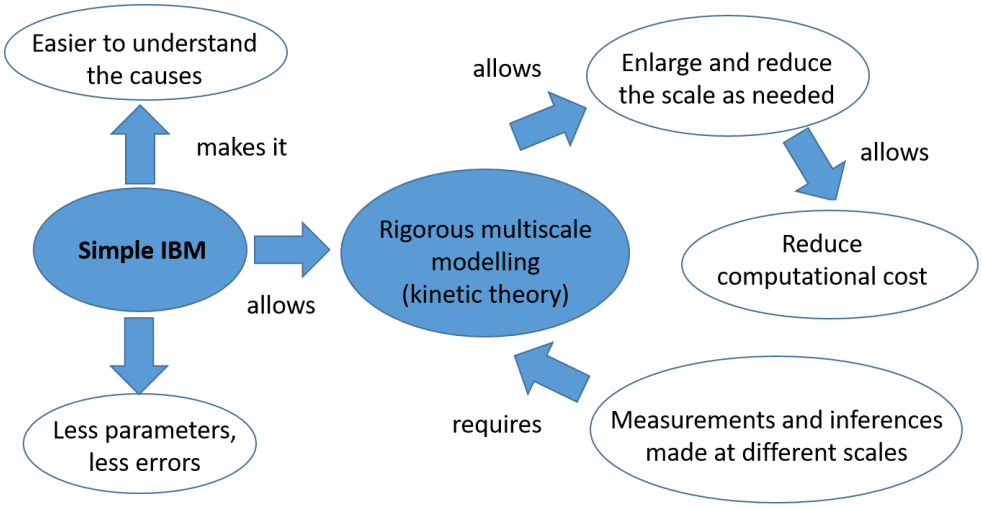


Figure 4.3: Scheme of our modelling strategy.

A fairly big amount of individual based models of packed tissues has recently been proposed by several authors. A few examples are cellular automata [115], Potts model [163], vertex model [59, 168, 65, 170], Voronoi model [127], subcellular model [159]. All these models use a complex representation of a cell. Simpler representations use quadrilaterals [91] and hexagonal prisms [84], however, they do not account for the variability in cell size or for the movement of each individual cell. In this work, instead, we use soft-hard-spheres, modelling the cell nuclei, and springs, modelling the cytoplasm, as presented in the next Section. The cells may differ from each other with respect to the rest length of the springs and to the radius of the nucleus.

4.3 Tissue at mechanical equilibrium

4.3.1 Model description

In this Section we first describe the model of one cell and the link to neighbouring cells. This is a 2D model representing a subsection of the neuroepithelium as presented in Figure 4.1c). Since we are primarily interested in studying the morphology of a cross-section of the neural tube and in particular the effect of INM, a 2D model is a reasonable starting point. However, a more accurate 3D model should be developed in the future in order to incorporate 3D effects, such as crowding in 3D and a 2D apical network (see Section 4.7.2 for a brief description of such extension).

A cell is constituted by the nucleus, an apical point and a basal point (see Figure 4.4). The nucleus is modelled by an inner-circle, which represents an impenetrable hard-core and an outer-circle, which represents a soft-core that resists to compression through an elastic response. The apical point represents the point of connection with neighbouring cells and the basal point represents the point where the cell is attached to the basement membrane. The two points are connected to the nucleus through springs with adaptive rest length, representing the cytoplasm. The rest length is gradually adapting to the actual length of the spring, which introduces a visco-elastic effect on the system. Each apical point is connected through springs to one or two neighbouring apical points, depending if the cell is in the edge or in the middle of the tissue, respectively. The apical connections represent the so-called apical network. Let N be the number of cells in the tissue. For each $i = 1, ..., N$ the parameters that characterize the configuration of the tissue are the following:

- the radii hard-cores of the nucleus $R_i^H > 0$,
- the position of the center of the nucleus, $X_i = (x_i(1), x_i(2)) \in \mathbb{R}^2$,
- the position of the point where the cell attaches to the basement membrane, $b_i = (b_i(1), b_i(2)) \in \mathbb{R}^2$,

- the position of the point where the cell attaches to neighbouring cells, $a_i = (a_i(1), a_i(2)) \in \mathbb{R}^2$,

For each $i = 1, ..., N$ and $j = 1, ..., N-1$ the springs, the alignment force and the nuclei soft-core are characterized by:

- the rest length η_i^{aX} and stiffness α_i^{aX} of the spring connecting the nucleus and the apical point
- the rest length η_i^{bX} and stiffness α_i^{bX} of the spring connecting the nucleus and the basal point,
- the rest length a_0 and stiffness α_j^a of the spring connecting apical points j and $j+1$,
- the strength of the alignment force, α_i^a,
- the stiffness of the soft nucleus, α_i^X.

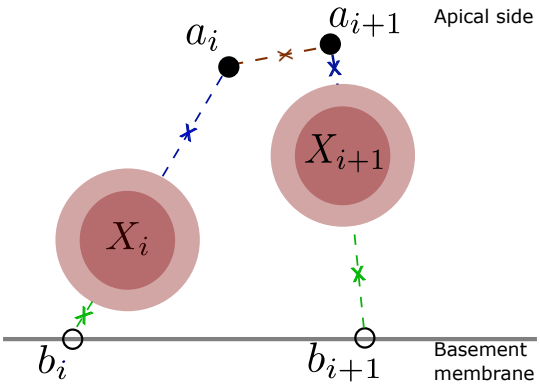


Figure 4.4: Sketch of the model for the tissue presented in Figure 4.2. The model consists of soft-hard-spheres connected to the apical and basal points through springs. The basal points are restricted to a straight line and they cannot switch positions nor get too far away from each other. Neighbouring apical points are connected through springs. The apical point, nucleus and basal point are subject to an alignment force.

Additionally, we consider the following rules. The basal points are restricted to the basement membrane, which is represented by a straight line, i.e., $b_i(2)$ is kept constant. Neighbouring basal points cannot switch positions nor get too far away from each other, i.e., $0 < b_{i+1}(1) - b_i(1) < 2\bar{R}b_0,\ i = 1, ..., N-1$, where $\bar{R} = \sum_i R_i^S/N$ and $b_0 > 0$. The rule that the basal points cannot be too far away from each other models lateral adhesion, i.e., the fact that cells adhere to each other. We also consider an alignment force that tends to align the apical point, the nucleus and the basal point, which models the influence of the cytoplasm of the cell.

4.3.2 Minimization problem

We describe the tissue at equilibrium as a local solution to a minimization problem for the positions of the nuclei, apical points and basal points. All the other parameters are

constants that are determined by the dynamics of the system, as explained in the next Section. We minimize a potential corresponding to the sum of the energy associated to the springs, the alignment force and the soft radius. The minimization is subject to non-overlapping on the nuclei inner-cores and non-switching and lateral adhesion constraints on the basal points.

Let $\mathbf{X} = (X_1, ..., X_N)$, $\mathbf{a} = (a_1, ..., a_N)$ and $\mathbf{b} = (b_1, ..., b_N)$ be the positions of the center of the nuclei, the apical points and the basal points, respectively. Consider the parameters defined in the previous Section. We define the potential $W : (\mathbb{R}^{2N})^3 \to \mathbb{R}$ by

$$W = W_1 + W_2 + W_3 + W_4 + W_5$$

with

$$\begin{aligned}
W_1(\mathbf{X}, \mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{N} \alpha_i^{aX} \left| \frac{|a_i - X_i|}{R_i^S(1 + \eta_i^{bX})} - 1 \right|^2 \\
W_2(\mathbf{X}, \mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{N} \alpha_i^{bX} \left| \frac{|b_i - X_i|}{R_i^S(1 + \eta_i^{bX})} - 1 \right|^2 \\
W_3(\mathbf{X}, \mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{N-1} \alpha_i^{a} \left| \frac{|a_{i+1} - a_i|}{2\bar{R}} - a_0 \right|^2 \\
W_4(\mathbf{X}, \mathbf{a}, \mathbf{b}) &= \sum_{i=1}^{N} \alpha_i^{ab} \left(\frac{(a_i - X_i) \cdot (b_i - X_i)}{|a_i - X_i||b_i - X_i|} + 1 \right)^2 \\
W_5(\mathbf{X}, \mathbf{a}, \mathbf{b}) &= \sum_{i,j=1}^{N} \alpha_{ij}^{X} \mathbb{1}_{\{|X_i - X_j| < R_i^S + R_j^S\}}(\mathbf{X}) \left| \frac{|X_i - X_j|}{R_i^S + R_j^S} - 1 \right|^2,
\end{aligned}$$

The potentials W_1, W_2 and W_3 represent the energy associated to the springs, W_4 represents the energy associated to the alignment forces and W_5 represents the energy associated to the soft-core of the nuclei.

Consider the functions associated to the constraints:

- $\phi_{k\ell}(\mathbf{X}) = \frac{(R_k^H + R_\ell^H)^2 - |X_k - X_\ell|^2}{\bar{R}^2}$, $k, \ell = 1, ..., N$, $k < \ell$,
- $\psi_i(\mathbf{b}) = \frac{b_i(1) - b_{i+1}(1)}{\bar{R}}$, $i = 1, ..., N-1$,
- $\Omega_i(\mathbf{b}) = \frac{b_{i+1}(1) - b_i(1)}{\bar{R}} - 2b_0$, $i = 1, ..., N-1$,

For $k, \ell = 1, ..., N$, $k < \ell$ and $i = 1, ..., N-1$, the non-overlapping, the non-switching and the lateral adhesion constraints are then respectively described by

$$\phi_{k\ell}(\mathbf{X}) \leq 0, \quad \psi_i(\mathbf{b}) \leq 0, \quad \Omega_i(\mathbf{b}) \leq 0.$$

The minimization problem is formulated as follows: find $(\bar{\mathbf{X}}, \bar{\mathbf{a}}, \bar{\mathbf{b}})$ such that

$$(\bar{\mathbf{X}}, \bar{\mathbf{a}}, \bar{\mathbf{b}}) \in \underset{\phi_{kl}(\mathbf{X}), \psi_i(\mathbf{b}), \Omega_i(\mathbf{b}) \leq 0,\ k<\ell}{\operatorname{argmin}} W(\mathbf{X}, \mathbf{a}, \mathbf{b}) \tag{4.1}$$

We use the damped Arrow-Hurwicz algorithm (DAHA) [45] developed in Part I to solve the minimization problem (4.1). We consider the Lagrangian $\mathcal{L} : (\mathbb{R}^{2N})^3 \times \mathbb{R}_+^{N(N-1)/2} \times (\mathbb{R}_+^{N-1})^2 \to \mathbb{R}$ associated to the potential W and to the constraints, which is defined by

$$\mathcal{L}(\mathbf{X}, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\xi}) = W(\mathbf{X}, \mathbf{a}, \mathbf{b}) + \sum_{k,\ell=1,\dots,N,\ k<\ell} \lambda_{k\ell}\phi_{k\ell}(\mathbf{X}) + \sum_{i=1}^{N-1} \mu_i \psi_i(\mathbf{b}) + \sum_{i=1}^{N-1} \xi_i \Omega_i(\mathbf{b})$$

where $\boldsymbol{\lambda} = \{\lambda_{k\ell}\}_{k,\ell=1,\dots,N,\ k<\ell}$, $\boldsymbol{\mu} = \{\mu_i\}_{i=1,\dots,N-1}$ and $\boldsymbol{\xi} = \{\xi_i\}_{i=1,\dots,N-1}$ are the Lagrange multipliers associated to the constraints. The DAHA consists on a discretization of the damped Arrow-Hurwicz system, that is given by

$$\begin{cases}
\ddot{X}_i &= -c\dot{X}_i - \alpha^2 \nabla_{X_i}\mathcal{L} - \gamma^2 \sum\limits_{k,\ell=1,\dots,N,\ k<\ell} \lambda_{k\ell}\phi_{k\ell}\nabla_{X_i}\phi_{k\ell} \\
\ddot{a}_i &= -c\dot{a}_i - \alpha^2 \nabla_{a_i}\mathcal{L} \\
\ddot{b}_i &= -c\dot{b}_i - \rho^2\alpha^2 \nabla_{b_i}\mathcal{L} - \gamma^2 \sum\limits_{k=1}^{N-1} \left[\mu_k \psi_k \nabla_{b_i}\psi_k + \xi_k \Omega_k \nabla_{b_i}\Omega_k\right], \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \dots, N \\
\dot{\lambda}_{k\ell} &= \begin{cases} 0, & \text{if } \lambda_{k\ell} = 0 \text{ and } \phi_{k\ell} < 0 \\ \beta\phi_{k\ell}, & \text{otherwise} \end{cases}, \ k, \ell = 1, \dots, N,\ k < \ell \\
\dot{\mu}_k &= \begin{cases} 0, & \text{if } \mu_k = 0 \text{ and } \psi_k < 0 \\ \beta\psi_k, & \text{otherwise} \end{cases}, \ k = 1, \dots, N-1 \\
\dot{\xi}_k &= \begin{cases} 0, & \text{if } \xi_k = 0 \text{ and } \Omega_k < 0 \\ \beta\Omega_k, & \text{otherwise} \end{cases}, \ k = 1, \dots, N-1
\end{cases} \tag{4.2}$$

where $\alpha,\ \rho,\ \beta$ and c are positive constants and $\rho \leq 1$. The parameter ρ in the equation for b_i models basal adhesion: the smaller it is, the slower the basal points move during the minimization algorithm. Consequently, due to the non-uniqueness of steady state solutions, the algorithm should reach a steady state where the basal points have not moved much from their initial positions, when compared to the nuclei and apical points. By discretizing this system, we get the DAHA, with the positive numerical parameters $\alpha,\ \beta,\ \gamma$ and c the tolerance associated to the stopping criterion, ϵ. The values used in the simulations are $\alpha = 0.01\bar{R},\ \beta = 0.25,\ \gamma = 0.1\bar{R},\ c = 2$ and $\epsilon = 10^{-6}$.

4.4 Dynamics driven by noise, cell division and changes in cell characteristics

4.4.1 Model description

In this section we describe the time-dependant model for the dynamics and growth of the tissue until either a maximum number of cells or a maximum time has been reached. The movement of cells is induced by cell division, changes in cell characteristics and cell diffusion with diffusion coefficient d.

We consider the cell cycle divided into three stages: S/G1, G2 and mitosis. We assume that a cell has an internal clock that determines the position of the cell in the cell cycle.

During the first stage S/G1, a cell moves passively driven by the crowd and noise, while trying to reach a internal state of equilibrium. This equilibrium is associated to the desired cell shape as explained in the previous Section. However, in order to accommodate the volume of nuclei, the tissue grows in all directions and in particular, in the apico-basal direction. Consequently, the cells have to stretch, i.e., the springs connecting the nucleus to the apical and basal points have to stretch. This leads to tension within each cell. We assume that cells release this tension, by letting the rest length of the springs progressively adapt to the actual size of the spring. This models the dynamics of actomyosin and microtubules inside a cell. Given the configuration of a cell at some time-point t, $(X_i(t), a_i(t), b_i(t))$, the desired rest lengths are given by $\eta_i^{aX*}(t) = |X_i(t) - a_i(t)|/R_i^S - 1$ and $\eta_i^{bX*}(t) = |X_i(t) - b_i(t)|/R_i^S - 1$. The dynamics of the rest lengths $\eta_i^{aX}(t)$ and $\eta_i^{bX}(t)$ are then governed by the piecewise linear ODE

$$\dot{\eta}_i^{aX} = \begin{cases} 0, & \text{if } \eta_i^{aX} = 0 \\ k_\eta(\eta_i^{aX*} - \eta_i^{aX}), & \text{otherwise} \end{cases}, \ i = 1, ..., N \tag{4.3}$$

and

$$\dot{\eta}_i^{bX} = \begin{cases} 0, & \text{if } \eta_i^{bX} = 0 \\ k_\eta(\eta_i^{bX*} - \eta_i^{bX}), & \text{otherwise} \end{cases}, \ i = 1, ..., N \tag{4.4}$$

where k_η is a positive constant controlling the speed of actualization of the rest lengths.

During stage G2, the characteristics of the springs change in order to encourage the nucleus to approach the apical side of the tissue:

- the spring connecting the neighbouring apical points of the dividing cell gets stronger (increase in α_i^a) and
- a pushing force from the basal membrane and a pulling force from the apical network are activated (increase of rest length of the spring connecting the nucleus to the basal membrane, η_i^{bX*}, and decrease of rest length of the spring connecting the nucleus to

the apical point, η_i^{aX*}),

- the spring connecting the nucleus to the apical point gets stronger (increase in α_i^{aX}).

During mitosis, besides the changes implemented during G2, two additional changes related to cell rigidity occur:

- the nucleus gets larger and more rigid (increase in R_i^H),
- cell gets more rigid (increase in alignment force in α_i^{ab}).

The cell will divide with probability P_{out} as soon as it reaches its maximum life time, despite the nucleus being or not at the apical side. The cell division is performed according to the following rules (see Figure 4.5):

- the division plan is parallel to the basement membrane: the nucleus, apical and basal points are substituted by two nuclei and two apical and basal points which are disposed near each other
- the two new cells and nuclei have the normal rigidity and size
- the springs of each cell and apical points are restored.

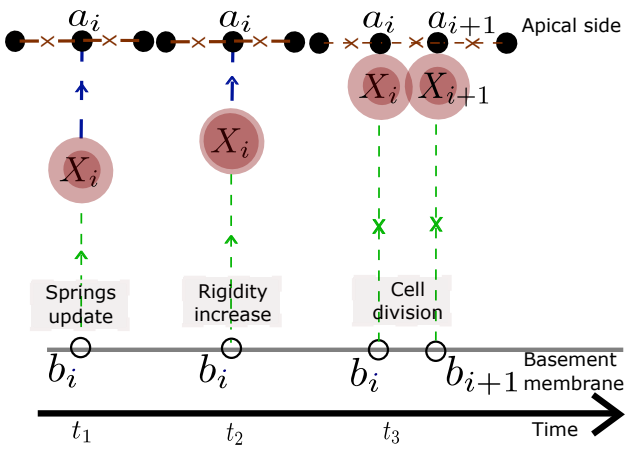


Figure 4.5: Sketch of the changes occurring to a cell prior to division. The rest length of the springs changes to allow the nucleus to approach the apical side. After a while, the radius of the hard nucleus increases and finally, at division, one cell is substituted by two cells.

Over the whole dynamics, the nuclei diffuse with coefficient of diffusion d, i.e., for each cell i, the vector X_i follows the equation $dX_i(t) = \sqrt{2d}\, dW(t)$, where $W(t)$ is the two-dimensional standard Wiener process.

For each cell i, we denote the duration of the $G2$ phase and mitosis by σ_i^{G2} and σ_i^M, respectively. Moreover, we denote the life time by σ_i and the birth time by τ_i. The time unit is hour. We use bold notation to denote the parameters for all cells $i = 1, ..., N$. We distinguish four types of parameters:

- tissue parameters: $a_0,\ b_0,\ \rho$ (defined in Section 4.3.1), $d,\ P_{\text{out}},\ k_\eta,\ N$
- cell cycle parameters: $\boldsymbol{\sigma}$, $\boldsymbol{\sigma}^{G2}$, $\boldsymbol{\sigma}^{M}$ and $\boldsymbol{\tau}$
- cell parameters: $\mathbf{R}^S$, $\mathbf{R}^H$, $\boldsymbol{\alpha}^{aX}$, $\boldsymbol{\alpha}^{bX}$, $\boldsymbol{\alpha}^{a}$, $\boldsymbol{\alpha}^{ab}$, $\boldsymbol{\alpha}^{X}$
- cell variables: $\mathbf{X}, \mathbf{a}, \mathbf{b}, \boldsymbol{\eta}^{aX}, \boldsymbol{\eta}^{bX}$

4.4.2 Time-stepping scheme

In this Section we present the time-stepping algorithm we use to simulate the dynamics described in the previous Section.

Let Δt be the time-step, which in practice will be equal to the time interval between two frames in the videos obtained in the lab, i.e., 0.1 hours. Let the initial tissue parameters, cell cycle parameters and cell parameters (see end of Section 4.4.1) be given. Consider the initial time $t_0 = 0$ and let the values for the cell variables $\mathbf{X}_0, \mathbf{a}_0, \mathbf{b}_0$, $\boldsymbol{\eta}_0^{aX}$ and $\boldsymbol{\eta}_0^{bX}$ be given. Given the cell variables at time t_{n-1}, we obtain $\mathbf{X}_n, \mathbf{a}_n, \mathbf{b}_n$ $\boldsymbol{\eta}_n^{aX}$ and $\boldsymbol{\eta}_n^{bX}$ at time $t_n = n\Delta t$ by performing the following steps:

1. evolve the rest length of springs of all cells i, i.e., obtain $\eta_{i,n}^{aX}$ and $\eta_{i,n}^{bX}$ through a time-discretization of equations (4.3) and (4.4) with $\eta_k^{aX*} = |X_{i,n-1} - a_{i,n-1}|/R_i^S - 1$ and $\eta_n^{bX*} = |X_{i,n-1} - b_{i,n-1}|/R_i^S - 1$ if cell i is in S/G1 phase and $\eta_n^{aX*} = 0$ and $\eta_n^{bX*} = |a_{i,n-1} - b_{i,n-1}|/R_i^S - 2$ if cell i is in G2 or mitosis.

2. identify the cells k that are going to enter in G2 phase and update the stiffness of the springs: increase $\alpha_k^{aX}, \alpha_{k-1}^{a}, \alpha_k^{a}$.

3. identify the cells k that are going to enter in mitosis and update the cell rigidity: increase R_k^H, and α_k^{ab}.

4. identify cells k that have reached the life time and divide them: increment N to $N+1$, remove the old cell and initialize two new cells k_1 and k_2 by performing the following steps:

   (a) restore the hard-core of the nucleus and the magnitude of the forces associated to the springs and alignment.

   (b) define the parameters: $\tau_{k_1} = \tau_{k_2} = t_n,\ \sigma_{k_1} = u_{k_1}$ and $\sigma_{k_2} = u_{k_2}$, where u_{k_1}, u_{k_2} are two values generated from the uniform distribution with support $[\sigma_{\min}, \sigma_{\max}]$ with $\sigma_{\min}, \sigma_{\max} > 0$,

   (c) actualize the variables:

$$\eta_{j,n}^{aX} = \eta_{j,n-1}^{aX},\ \eta_{j,n}^{bX} = \eta_{j,n-1}^{bX},\ j = k_1, k_2$$

$$\hat{\hat{X}}_{k_1,n} = X_{k_1,n-1} - (0.05R_k^S, 0), \quad \hat{\hat{X}}_{k_2,n} = X_{k_2,n-1} + (0.05R_k^S, 0)$$

$$\hat{a}_{k_1,n} = a_{k_1,n-1} - (0.05R_k^S, 0), \quad \hat{a}_{k_2,n} = a_{k_2,n-1} + (0.05R_k^S, 0)$$
$$\hat{b}_{k_1,n} = b_{k_1,n-1} - (0.05R_k^S, 0), \quad \hat{b}_{k_2,n} = b_{k_2,n-1} + (0.05R_k^S, 0),$$

5. obtain $\hat{\mathbf{X}}_n$ by adding to $\hat{\hat{\mathbf{X}}}$ Gaussian white noise with mean 0 and standard deviation $\sqrt{2d\Delta t}$, i.e., $\hat{\mathbf{X}}_n = \hat{\hat{\mathbf{X}}}_n + \sqrt{2d\Delta t}u$, where u is generated from the $2N$-dimensional multivariate standard Gaussian distribution,

6. obtain an admissible configuration $(\mathbf{X}_n, \mathbf{a}_n, \mathbf{b}_n)$ by readjusting the positions in $(\hat{\mathbf{X}}_n, \hat{\mathbf{a}}_n, \hat{\mathbf{b}}_n)$. In other words, obtain a local solution, $(\mathbf{X}_n, \mathbf{a}_n, \mathbf{b}_n)$, of (4.1) by numerically solving the ODE system (4.2) with initial condition $(\hat{\mathbf{X}}_n, \hat{\mathbf{a}}_n, \hat{\mathbf{b}}_n)$.

The algorithm stops when either the maximum number of cells or the maximum time have been reached.

4.5 Numerical results

In Sectio 4.5.1 we present the ranges of the parameters and the type of initial condition that are used in the remaining Sections of the Chapter. In Section 4.5.2 we study the effect of the magnitude of the forces associated to the springs, alignment and soft nucleus. In Section 4.5.3 we study the effect of noise. Finally in Section 4.5.4 we study the evolution of the tissue over 36 hours and we compare the results with experimental data.

4.5.1 Choosing the model parameters and initial conditions

The ranges for the model parameters are presented in Table 4.1. These ranges are chosen according to the data obtained from lab experiments conducted by Theveneau's lab or taken from the literature. The aim is to reproduce a section of the neuroepithelium as shown in Figure 4.2. Theveneau's lab has obtained statistical descriptors of the tissue shape and morphology for different stages of development. We do not present these descriptors here as we rather keep the focus on the model. We only mention some averaged values that allow us to get a first estimate of the model parameters. In the future, more accurate estimates should be obtained. We also resort to image 4.2 to justify some choices for some parameters. Despite being just one single sample, the lab experiments have shown that there is a relatively low variability among different embryos. Thus the tissue represented in image 4.2 can be considered as a good and robust representation of the system. Some parameters will be obtained by direct measurements on that image.

We start by choosing the initial number of cells $N = 30$ as the minimum number that allow us to reproduce the morphology of the tissue.

We define the radius of the soft-core of the nucleus as the space unit and we use the value $R_i^S = 1$ for all i. Measurements on the shape of the nucleus indicate that the shape

Name	Symbol	Unit	Range		
Tissue parameters					
Number of cells	N		$[30, 60]$		
Speed of actualization of the rest length of nucleus springs	k_η	$1/h$	≥ 2		
Rest length of apical springs	a_0	$2\bar{R}$	$\leq 1/6$		
Maximal distance between neighbouring basal points	b_0	$2\bar{R}$	$[1/3, 1/2]$		
Minimizer selector	ρ		$1/2$		
Diffusion coefficient	d	$\bar{R}^2/h$	2.5		
Proportion of outer divisions	P_{out}		0.8		
Cell cycle parameters					
Time life	σ	h	$[11, 14]$		
Duration of G2 phase	σ^{G2}	h	$1/2$		
Duration of M phase	σ^{M}	h	$1/2$		
Cell parameters					
			S/G1	**G2/M**	**M**
Radius of soft-core of the nucleus	R^S		1		
Radius of hard-core of the nucleus	R^H		$R^{H_1} \in [R^S/2, R^S]$		$R^{H_2} \in [R^{H_1}, R^S]$
Stiffness of spring connecting nucleus and apical point	α^{aX}		$\mathcal{O}(1)$	$\mathcal{O}(1)$	
Stiffness of spring connecting nucleus and basal point	α^{bX}		$\mathcal{O}(1)$	$\mathcal{O}(1)$	
Stiffness of spring connecting neighbouring apical points	α^{a}		$\mathcal{O}(10)$	$\mathcal{O}(10)$	
Magnitude of alignment force	α^{ab}		$\mathcal{O}(10)$		$\mathcal{O}(10)$
Stiffness of soft-core of the nucleus	α^{X}		$\mathcal{O}(1)$	$\mathcal{O}(1)$	

Table 4.1: Ranges of model parameters obtained as much as possible from lab experiments. The symbol $\mathcal{O}$ denotes order of magnitude and $\bar{R} = \sum_i R_i^S/N$.

is close to an ellipsoid with axis a, b, c, where $a = b < c$ and $c = 1.5a$ or $c = 2a$. To get a close representation we consider that the larger axis corresponds to the radius of the soft-core and we choose the value for the radius of the hard-core within the longer and shorter axis, $R_i^H \in [R_i^S/2, R_i^S]$. During mitosis the value radius the radius of the hard-core is larger than outside mitosis.

The coefficients representing the stiffness of the springs do not affect directly the dynamics. Instead they have an direct role in the minimization algorithm. Since there are many local minima, these coefficients in some sense impose a priority on which energy associated to each force to minimize first. Consequently, the first one to be minimized will be closer to the minimum than the others, thus representing a stiffer spring or a stronger force. This allows us to choose the parameters roughly and relatively to each other. We know from the experiments that both the springs between apical points and the alignment force, should be much stronger than the other forces. So we choose $\alpha^X, \alpha^{aX}, \alpha^{bX}$ of order 1 and α^{ab}, α^a of order 10. In the next Section we show the effect of considering other values for these parameters. During G2 and mitosis the stiffness of the soft nucleus and springs is larger than outside these stages.

We observe in the experiments that the nucleus starts to approach the apical side 1 hour before division and that within half an hour it arrives there. To reproduce this behaviour, the rest length of the springs connected to the nucleus have to actualize in less than half an hour. Recall that, during this stage, the desired rest lengths are 0 and $|a_i - b_i|$ for the springs connecting the nucleus and the apical and basal points, respectively. Since k_η corresponds to the inverse of the time it takes for η_{aX} and η_{bX} to reach such value, we have to choose k_η larger than the inverse of half an hour, i.e., $k_\eta \geq 2/h$, where h denotes hour. At the beginning of the dynamics, the rest lengths are initialized with the value 3.

The quantity $2a_0\bar{R}$, with $\bar{R} = \sum_i R_i^S/N$ corresponds to the rest length between apical points. From the experiments, we observe that the apical network is approximately straight. Therefore, the distance between apical points i and $i+1$ has to be approximately equal to the quotient between $2R_i^S$ and the number of nuclei that lie within the rectangle with width $2R_i^S$ and length equal to the height of the tissue, i.e., the nuclei that lie exactly below or above nucleus i. This number of nuclei is equal to the number of layers of nuclei in the tissue. At the initial stage, we observe around 6 to 8 layers of nuclei. So we have $2a_0\bar{R} = 2\bar{R}/6$ and therefore $a_0 = 1/6$.

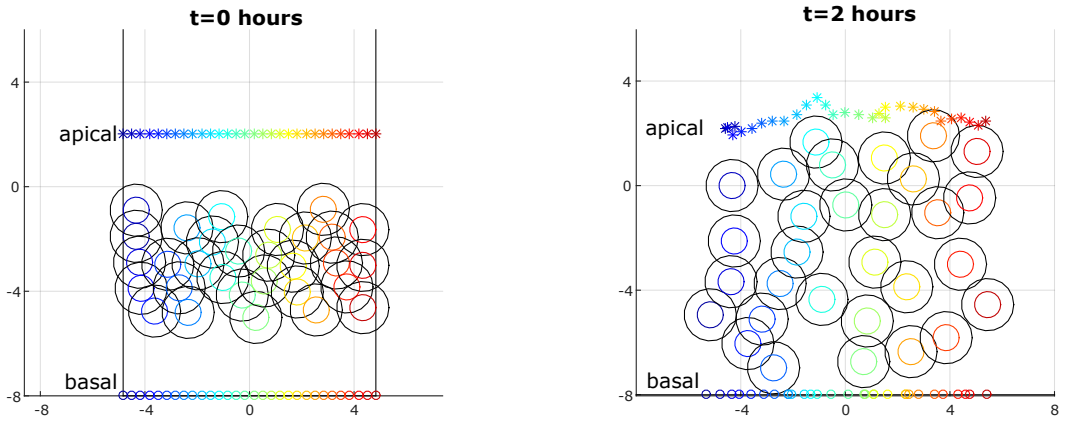
The parameters $\rho \leq 1$ and b_0, are related to basal adhesion. The parameter ρ contributes to slow down the dynamics of the basal points in the minimization algorithm and consequently, select a minimizer in which the basal points have not moved much from their previous positions. We use the value $\rho = 1/2$. The quantity $2b_0\bar{R}$ corresponds to the maximal distance between basal points. We will consider $b_0 \in [1/3, 1/2]$. Note that when the tissue has a rectangular shape, the distance between basal points should be similar to the distance between apical points. As mentioned in the previous paragraph, such distance should be around $2\bar{R}/6$. Consequently, in this case, the bound imposed on the basal points is not limiting their movement too much.

The diffusion coefficient d has been estimated from the movement of nuclei *in vivo* as follows. Given the successive positions of a nucleus over the life time of the cell, we selected a time window with 15 frames in which the nucleus is shaking without having any apparent directed movement. We measure the consecutive displacements $\Delta x_n,\ n = 1, ..., 14$ during this time window and we obtain the mean $\hat{d}$ and standard deviation $\hat{\sigma}$ of the values $d_n = (\Delta x)_n^2/(2\Delta t)$, where $\Delta t = 0.1$ hours. We did this first for the data in 3D. Then we ignored the third coordinate which gives the longitudinal position in the neural tube and we obtained an estimate of the diffusion coefficient in 2D. In the 3D case, we obtained $\hat{d} = 0.37\bar{R}^2/h$ and $\hat{\sigma} = 0.61\bar{R}^2/h$ and in 2D $\hat{d} = 0.32\bar{R}^2/h$ and $\hat{\sigma} = 0.60\bar{R}^2/h$. Naturally the estimate in 3D is larger than in 2D, as there is the movement in the third dimension that we are ignoring in the 2D estimate. In fact, since the model is in 2D, the nuclei have less space to move than in the real tissue. This may cause extra difficulties to the overall dynamics of the tissue that do not exist *in vivo*. One way to overcome this limitation of the model is by considering a larger diffusion coefficient. On the other hand, if we look at the estimated standard deviation, we see that its value is quite large, which is due to

a few big jumps with an amplitude that can reach the order of magnitude of the radius within one hour. This shows that the diffusive behaviour of the nucleus is very complex and should be further statistically explored. As a first rough approximation, we then consider a diffusion coefficient of $d = 2.5\bar{R}^2/h$.

The duration of the cell cycle ranges approximately between 12 and 14 hours and mitosis and G2-phase[§] take around 30 minutes [108]. In the model, we generate the duration of the cell cycle σ_i for each cell i from a uniform distribution $\mathcal{U}([11, 14])$. We use a smaller lower bound because some cells might take less than 12 hours. The duration of G2 and mitosis are always the same for all cells and they are given by $\sigma^{G2} = 1/2$ and $\sigma^M = 1/2$, respectively.

The divisions are observed to occur in the plan perpendicular to the apico-basal axis. This means that when a cell divides, one daughter cell may be formed outside the cross section we are looking at. We assume that this happens when the direction of division forms an angle larger than 18° with the plan defined by the cross section. This value corresponds to 20% of all possible angles. Assuming that the divisions can occur in any direction with equal probability within the plan of division mentioned before, we conclude that the probability that a division occurs outside the plan of the cross section is $P_{\text{out}} = 0.8$.



(a) Initial condition.

(b) Tissue after 2 hours.

Figure 4.6: Initial (a) and final (b) tissue configuration obtained by the algorithm described in Section 4.4.2. The stars represent apical points, the small circles at the basal line represent the basal points, the coloured circles represent the hard-core of the nucleus and the black circles represent the soft-core. The tissue parameters are $N = 30$, $a_0 = 1/6$, $b_0 = 1/3$, $\rho = 1/2$, $d = 2.5$, $P_{\text{out}} = 0.8$. The cell cycle parameters are $\sigma \in [11, 14]$, $\sigma^{G2} = 1/2$ and $\sigma^M = 1/2$. The cell parameters during G1/S are $R^S = 1$, $R^H = 0.5$, $\alpha^{aX} = 2$, $\alpha^{bX} = 2$, $\alpha^a = 5$, $\alpha^{ab} = 15$, $\alpha^X = 1$. During G2/M we use $\alpha^a = 10$, $\alpha^{ab} = 15$ and during M we use $R^H = 0.8$.

In Figure 4.6a) we plot the initial configuration of the tissue $(\mathbf{X}_0, \mathbf{a}_0, \mathbf{b}_0)$ with $N = 30$ cells. The initial condition is obtained by defining the positions of the apical and basal

[§]To be precise, G2-phase takes about 2 hours and the nucleus moves apically during the last half an hour of this phase [108]. But for simplicity we call G2 to this period in which the nucleus migrates apically.

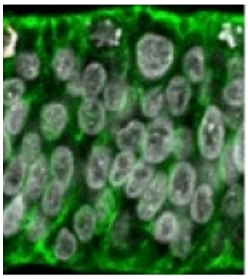
points along parallel horizontal line and randomly throwing nuclei in between these two lines. We then solve a minimization problem for the nuclei only, while keeping the apical and basal points fixed. During the minimization, the nuclei are restricted to two vertical walls, as we can see in Figure 4.6a). Each color represents one cell that is formed by the apical point, nucleus and basal point. The basal points are displayed in a horizontal line and the apical points at the top of the configuration. In the middle we see the nuclei with the soft-core (in black) and the hard-core (in color). For each cell i, we randomly generate the life time σ_i from the uniform distribution $\mathcal{U}([11, 14])$ and the birth hour τ_i from $\mathcal{U}([-\sigma_i, 0])$. The rest length of the springs connecting the nucleus and the apical and basal points for each cell i is initialized by $\eta_{i,0}^{aX} = \eta_{i,0}^{bX} = 3$.

In Figure 4.6b) we plot the configuration obtained after 2 hours for the parameters taken within the ranges mentioned above. We observe that the nuclei are more spread out and the tissue keeps a rectangular shape, despite not having lateral walls to hold the peripheral cells.

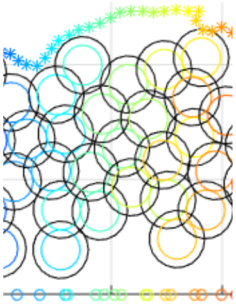
4.5.2 Influence of magnitude of the forces

In this Section we explore different values associated to the magnitude of the alignment and elastic forces presented in the system. This is controlled by the parameters $\alpha^X, \alpha^{aX}, \alpha^{bX}$ α^{ab} and α^a. Specifically, we study the effect of having $\alpha^X, \alpha^{aX}, \alpha^{bX}$ of order 10 and α^{ab}, α^a of order 1. The remaining parameters are chosen within the ranges presented in the previous Section (see Table 4.1). The experimental set up is the following. We start by generating an initial condition as described in the previous Section 4.5.1. We then let the system evolving for at least 2 hours and we plot the final configuration in Figure 4.7. We do this for one value of each parameter referred in the previous paragraph.

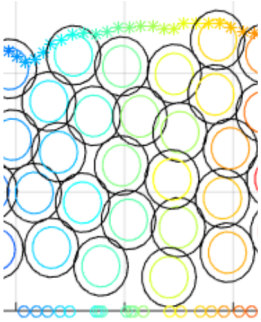
We first plot in Figure 4.7b) the control tissue that was obtained using parameters within the ranges presented in the previous Section. By comparing with the tissue obtained *in vivo* presented in Figure 4.7a), we observe that the overall configuration seems to coincide. In particular, the nuclei are displayed between the basal membrane and apical network, the number of layers of nuclei is around 6 to 8 and the relative size of the nucleus with respect to the size of the tissue seem pretty comparable. The apical membrane seems very straight in the real tissue and less straight in the simulated one. This maybe due to the fact that the apical network is a 2D network while in this model we represent it by a 1D network, which is in some sense less robust to the forces exerted by the nuclei (see Section 4.7.1 for improvements and extensions of the model). Moreover, the tiny green lines in between nuclei in Figure 4.7a) indicate that the cells are approximately straight, which is also observed in Figure 4.7b) by comparing the abscissa of basal point, nucleus and apical point of each cell. A more quantitative comparison between *in vivo* and *in silico* results is presented in Section 4.5.4.



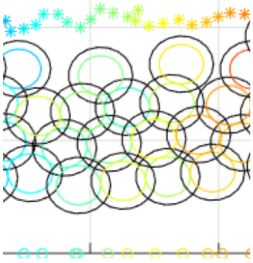
(a) Tissue obtained *in vivo*.



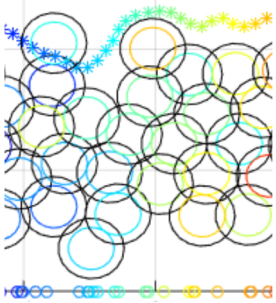
(b) Control tissue ($\alpha^X, \alpha^{aX}, \alpha^{bX} = 1$ and $\alpha^{ab}, \alpha^a = 10$).



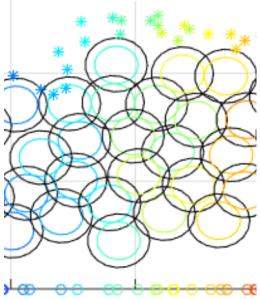
(c) Large stiffness of soft nucleus ($\alpha^X = 10$).



(d) Large stiffness of nucleus springs ($\alpha^{aX} = \alpha^{bX} = 10$).



(e) Weak alignment force ($\alpha^{ab} = 1$).



(f) Small stiffness of apical springs ($\alpha^a = 1$).

Figure 4.7: Final configurations obtained after 2 hours for different values of the parameters that control the stiffness of springs and nucleus and the magnitude of the alignment force. We observe different effects on the layering, packing, height of the tissue, straightness of each cell and of apical domain. The tissue parameters are $a_0 = 1/6$, $b_0 = 1/2$, $\rho = 1/2$, $d = 2.5$, $P_{\text{out}} = 0$. In this simulation the rest length of the springs is not evolving over time, $k_\eta = 0$, and they have the value $\eta_{aX} = \eta^{bX} = 3$ during G1/S and $\eta_{aX} = 1$, $\eta^{bX} = 10$ during G2/M. The cell cycle parameters are $\sigma = 10$, $\sigma^{G2} = 2$ and $\sigma^M = 1/2$. The remaining cell parameters during G1/S are $R^S = 1$ and $R^H = 0.7$. During G2/M we use $\alpha^a = 100$, $\alpha^{ab} = 20$ and during M $R^H = 0.9$. The initial condition is described in Section 4.5.1.

In Figure 4.7c) we consider the case of a more rigid soft nucleus, by using $\alpha^X = 10$. Comparing with the control tissue, we observe that the nuclei are more rigid and packed. This brings extra difficulties to the movement of nuclei, especially before mitosis and by looking at the video of the dynamics we see that many nuclei to not manage to reach the apical side before dividing.

In Figure 4.7d) we consider the case where the springs connecting the nucleus with the apical and basal points are stiffer by using $\alpha^{aX} = \alpha^{bX} = 10$. We observe less layers (around 4) and a shorter tissue.

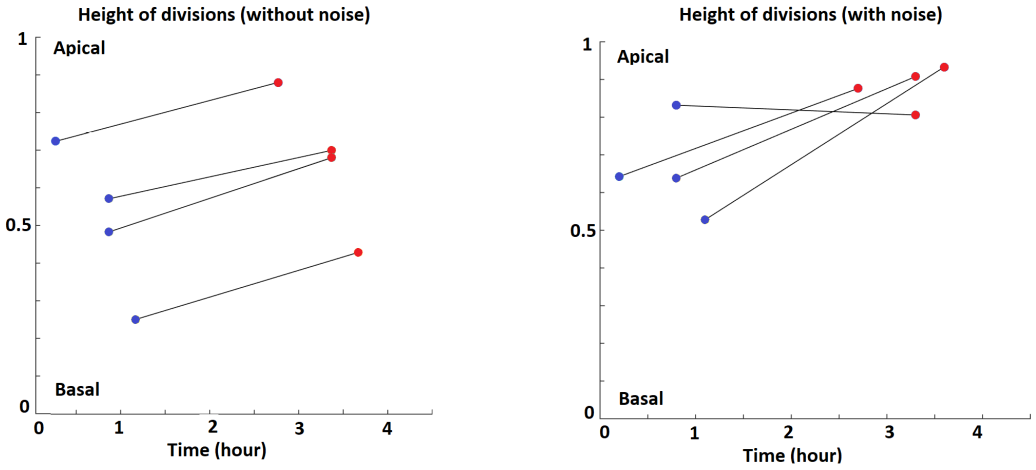
In Figure 4.7e) we consider a weaker alignment force by using $\alpha^{ab} = 1$. We observe a disorganization in the colors of the nuclei, indicating that the cells are not straight, in the sense that the basal point, the nucleus and the apical point are not aligned. Interestingly, if we look at the top left of the tissue we even see one nucleus above the apical domain, which typically does not happen in a healthy tissue, but may happen in a defected one.

Finally, in Figure 4.7f) we consider a small stiffness of apical springs by using $\alpha^a = 1$. We see that the apical domain is not straight and is quite rough. The tissue seems also

slightly shorter.

4.5.3 Influence of noise

In this Section we study the influence of noise on the positions of the nuclei and apical points during division along the apico-basal axis during a period of 4 hours. Specifically, we consider the vertical distance between the nucleus and the basal point normalized by the vertical distance between the apical point and the basal point, i.e., $|X_i(2) - b_i(2)|/|a_i(2) - b_i(2)|$. If the nucleus is near the apical point, then this quantity is close to 1. For each cell dividing within 4 hours, we plot this quantity for two time points: when the cell starts to approach the apical domain (in blue) and when the cell is dividing (in red). These points are connected to a straight line that indicates the global vertical direction of the nucleus during this period. We expect that this line has a positive slope and that the second dot lies at the top, near the apical domain. We obtain this plot



(a) $d = 0$. **(b)** $d = 2.5$.

Figure 4.8: Vertical position of the nucleus in the tissue normalized by the vertical distance between apical and basal point over time for each cell that has divided within 4 hours and for two time-points corresponding to the moment when the nucleus starts to go up (in blue) and when the cell divides (in red). We observe that in the case with noise (b) the nuclei divide apically, while in the case without noise (a) they don't. The tissue parameters are $a_0 = 1/6$, $b_0 = 1/2$, $\rho = 1/2$, $P_{\text{out}} = 0$. In this simulation the rest length of the springs is not evolving over time, $k_\eta = 0$, and they have the value $\eta_{aX} = \eta^{bX} = 3$ during G1/S and $\eta^{aX} = 1$, $\eta^{bX} = 10$ during G2/M. The cell cycle parameters are $\sigma = 10$, $\sigma^{G2} = 2$ and $\sigma^M = 1/2$. The cell parameters during G1/S are $R^S = 1$, $R^H = 0.7$, $\alpha^{aX} = 1$, $\alpha^{bX} = 1$, $\alpha^a = 10$, $\alpha^{ab} = 10$, $\alpha^X = 1$. During G2/M stage we use $\alpha^a = 100$, $\alpha^{ab} = 20$ and during M stage we use $R^H = 0.9$. The initial condition is described in Section 4.5.1.

for two values of the diffusion coefficient, namely, $d = 0$, corresponding to the situation without noise (see Figure 4.8a)) and $d = 2.5$, corresponding to the situation with noise (see Figure 4.8b)). We observe that 4 divisions have occurred within 4 hours. In the case without noise we observe that the nuclei approach the apical domain, however most of them do not manage to arrive there. In contrast, in the situation with noise, the nuclei approach the apical domain with a larger slope than in the previous case and most of

them manage to arrive there. These results suggest that the noise in the nuclei and apical points positions is fundamental to allow the nucleus migration to the apical side before division.

4.5.4 Evolution over 36 hours and comparison with experimental data

In this Section we consider the evolution of the tissue over 36 hours. We start by introducing quantifiers that provide information about the evolution of the shape and morphology of the tissue. Then we run simulations for one parameter set and five initial conditions and we plot the average and standard deviation for each quantifier over time. Some quantifiers have also been estimated in the lab by Theveneau's team using data from several embryos and several sections of the neuroepithelium from each embryo. We do not present all their results here, we simply consider the average value and we plot over the model results. This allow us to have a preliminary quantitative comparison with experimental data.

Let p be the number of initial conditions. The quantifiers used in this Section are presented next:

- *Number of layers*, N_{layers}. The number of layers is obtained by using a Voronoi tessellation [43] as follows. The Voronoi tessellation divides the space in regions containing exactly one nucleus each. We construct the tessellation in a way such that the nuclei situated at the top layer of the tissue belong to a region with infinite area, while the others don't. We compute recursively each layer by identifying the regions with infinite area and removing them from the tessellation.

- *Length of the apical domain, nuclei configuration and basal domain (unit: $\bar{R}$)*,

$$L_{\text{apical}} = a_N(1) - a_1(1),$$

$$L_{\text{nuclei}} = x_{\text{right}} - x_{\text{left}}$$

  and

$$L_{\text{basal}} = b_N(1) - b_1(1),$$

  where $x_{\text{left}} = \min_i X_i(1)$, $x_{\text{right}} = \max_i X_i(1)$.

- *Average height of the tissue and average height of nuclei (unit: $\bar{R}$)*,

$$H_{\text{tissue}} = \frac{1}{\#S_{\text{middle}}} \sum_{i \in S_{\text{middle}}} \frac{|a_i(2) - b_i(2)|}{R_i^S}$$

and

$$H_{\text{nuclei}} = \frac{1}{\#S_{\text{middle}}} \sum_{i \in S_{\text{middle}}} \frac{|X_i(2) - b_i(2)|}{R_i^S},$$

where $S_{\text{middle}} = \{i : x_{\text{left}} + L_{\text{nuclei}}/4 \leq X_i(1) < x_{\text{right}} - L_{\text{nuclei}}/4\}$ and $\#S_{\text{middle}}$ is the number of elements in S_{middle}.

- *Packing fraction of the hard-core of nuclei in the upper and lower half of the tissue,*

$$\Phi_{\text{up}} = \frac{2}{A_{\text{middle}}} \sum_{i \in S_{\text{up}}} \pi (R_i^H)^2$$

and

$$\Phi_{\text{down}} = \frac{2}{A_{\text{middle}}} \sum_{i \in S_{\text{down}}} \pi (R_i^H)^2,$$

with $S_{\text{up}} = \{i : \ (X_i(2) - b_i(2))/R_i^S > H_{\text{tissue}}/2\}$, $S_{\text{down}} = S_{\text{middle}} \backslash S_{\text{up}}$ and $A_{\text{middle}} = H_{\text{tissue}} \bar{R}(L_{\text{nuclei}}/2 + 2\bar{R})$, with $\bar{R} = \sum_i R_i^S/N$.

- *Straightness of apical network,*

$$A_{\text{straight}} = \left(\frac{1}{|a_1 - a_N|} \sum_{i=1}^{N-1} |a_i - a_{i+1}| \right)^{-1}$$

- *Height of nucleus i during division,*

$$H_{\text{div}} = \frac{|X_i^n(2) - b_i^n(2)|}{|a_i^n(2) - b_i^n(2)|}, \text{ with } t^n < \tau_i + \sigma_i < t^{n+1}.$$

We compute the value of each quantifier over time and for 5 different initial conditions. We obtain the mean and standard deviation for each time t_n and for each quantifier, except P_{div}. Since there may be times without any division, we compute the mean and standard deviation of P_{div} over a larger time-interval with size $5\Delta t$, which in this case corresponds to half an hour. In Figure 4.9 we plot the average and standard deviation for each quantifier over time $t \in [0, 36]$ for parameters taken within the ranges presented in Table 4.1.

We observe in Figure 4.9a) that the number of layers of nuclei grows approximately linearly from around 8 to 13 layers, which is in agreement with experimental data.

Figure 4.9c) shows the packing fraction of the nuclei hard-core at the top and at the bottom of the tissue. We observe that the packing fraction at the upper part remains approximately constant, while at the lower part it decreases. If we compare this result with the evolution of the height of nuclei presented in graph 4.9b) we conclude that the nuclei are moving away from the basement membrane. In graph 4.9b) we also observe that the height of the tissue is growing at a similar speed than the average height of nuclei. This growth is slightly slower when compared to experimental data.

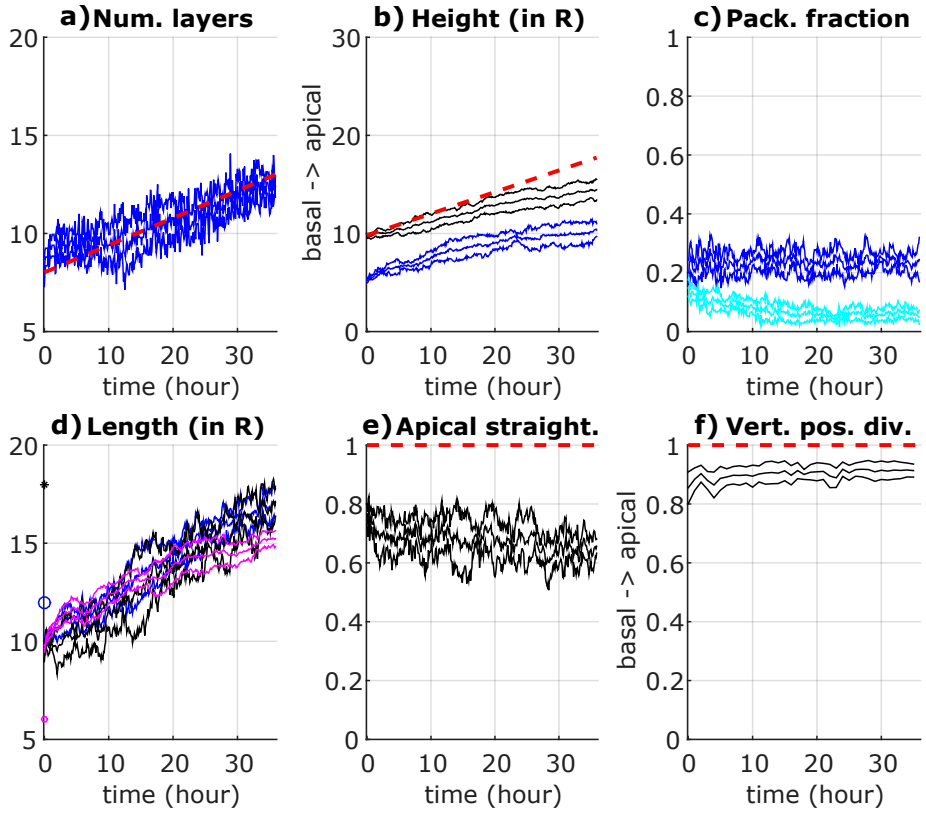


Figure 4.9: Statistical quantifiers describing the morphology of the tissue over 36 hours. For each graph we plot the mean and standard deviation over 5 initial conditions. When available we also plot the reference values obtained from experimental data (dashed red line). In Figure (a) we plot the number of layers of nuclei (blue), in Figure (b) the average height of nuclei (blue) and of the tissue (black), in Figure (c) the packing fraction of nuclei at the upper (dark blue) and lower (light blue) half of the tissue, in Figure (d) the length of the tissue using the basal points (pink), the nuclei (blue) and the apical points (black), in Figure (e) the apical straightness and in Figure (f) the height of nuclei during cell division. The tissue parameters are $N = [30, 53]$, $a_0 = 1/6$, $b_0 = 1/3$, $\rho = 1/2$, $d = 2.5$, $P_{\text{out}} = 0.8$. The cell cycle parameters are $\sigma \in [11, 14]$, $\sigma^{G2} = 1/2$ and $\sigma^M = 1/2$. The cell parameters during G1/S are $R^S = 1$, $R^H = 0.5$, $\alpha^{aX} = 2$, $\alpha^{bX} = 2$, $\alpha^a = 5$, $\alpha^{ab} = 15$, $\alpha^X = 1$. During G2/M we use $\alpha^a = 10$, $\alpha^{ab} = 15$ and during M we use $R^H = 0.8$.

Figure 4.9d) shows the evolution of the length of the tissue at the bottom, in the middle and at the top, by considering the length occupied by basal points, nuclei and apical points, respectively. Since the three quantities grow at approximately the same speed, we conclude that the shape of the tissue remains approximately rectangular, which is what we observe in experimental data.

Figure 4.9e) shows that the apical network is not very straight and that its straightness decreases over time. This result diverges from experimental results.

Figure 4.9f) shows the height of divisions occurring during every half an hour. We observe that the divisions occur apically, which is in agreement with experimental data.

In Figure 4.10 we show two samples of the tissue obtained from a 2-days old (a) and a 3.5-days old (b) embryos. We observe an increase of around 5 layers of nuclei. The height

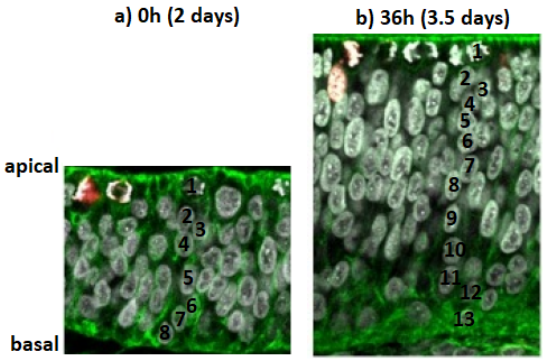


Figure 4.10: Neuroepithelium obtained from a 2-days old (a) and 3.5-days old (b) embryo. We observe that the older tissue is around 1.7 times taller and it has approximatelly more 5 layers of nuclei than the younger tissue. Images obtained by Theveneau's lab.

of the tissue is approximately 1.67 times higher in the older tissue compared with the younger one. Looking at the top of the images, we see that the apical network is almost completely straight in both cases.

4.6 Discussion

As we have seen in Section 4.5, the model has been able to reproduce many features of the shape and dynamics of the neuroepithelium. This may indicate that we have been able to identify the main mechanisms that underlie the dynamics of this system. We will therefore keep working along the biological assumptions regarding the way cells behave and interact with their neighbours and the basement membrane in order to obtain more robust results about their validity and to further explore the relation between crowding, INM, tissue shape and growth. At this point, the model has already suggested some new hypothesis that we present next.

The parameter k_η that controls the rest length of the springs connected to the nucleus seems to be a key parameter in the dynamics. Indeed, k_η seems to control directly both the height of the tissue and the speed of the nuclei during the apical movement before division. The speed should be slightly smaller than $|a_i - X_i|/k^\eta$ and the height of the tissue seems to grow with k_η. The interesting aspect, is that for a certain value of this parameter, we are able to get these two quantities close to what is observed in experimental data. This suggests that in reality the speed of the movement of the nuclei is controlling the height of the tissue.

This parameter k_η seems to have also an indirect role on the number of layers. The larger it is the larger the number of layers is. Note that a taller tissue does not necessarily imply a tissue with more layers, as the nuclei density is decreasing near the basement membrane over time. This might suggest that the lateral pressure between cells is not that high and that the packing is not playing an important role in the lateral direction. This

is supported also by the fact that in the model, despite of not having lateral compressive forces, the shape of the tissue is kept approximately rectangular.

The packing may instead have an important role in the upper part of the tissue near the apical membrane. Preliminary *in silico* experiments in the case of cells with a shorter life time (see Section 4.7.3) indicate that the nuclei is not able to reach the apical domain due to the too high packing fraction of nuclei near the domain. This is also observed in *vivo*. It would therefore be worth exploring the role of packing along the apico-basal axis.

Finally, we point out that the mechanisms we consider in the model are not specific to the neuroepithelium. Therefore our model could also be used to study other packed tissues. In Section 4.7.3 we present a possible application to the study of cancer tissue.

4.7 Future work

4.7.1 Model improvement, quantitative validation and parameter space exploration

The model has been able to reproduce the general aspects of the tissue regarding shape, morphology and the dynamics of cells before division. However there are a few aspects that can still be improved in order to get a representation that is closer to the data. One aspect where the model could be improved is on the straightness of the apical network. One possible way to increase the straightness could be by decreasing the rest length of the apical springs, $2\bar{R}a_0$, over time. This is indeed what happens in the real tissue as reported in [165]: as the tissue grows, the number of layers increases and the apical domain shrinks. Other modelling aspects that could be improved are described in the next Section 4.7.2.

In the future, a more quantitative comparison of the model with experimental data should be conducted. In particular, the comparison between the evolution of the packing fraction and tissue length are still missing.

Moreover, one can try to explore other parameter regimes, such as a larger number of cells N, a longer time period, different relations between the strength of the forces, different values for the radius of the soft and hard nucleus, etc. In particular, as referred in the discussion, the parameter k_η that controls the speed of actualization of the rest length of the springs should be further explored, as it may help to better understand the relation between cell behaviour and tissue shape and size. Output parameters, such as the packing density or the number of layers, could then be obtained for each value of input parameter, in order to look for bifurcations in the system. These bifurcations would then be searched *in vivo*. If they prove to exist, we would obtain a stronger validation of the model and a deeper understanding of the mechanics underlying the cell tissue.

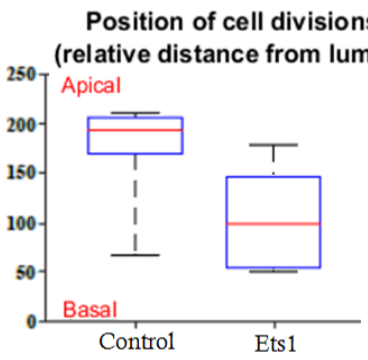
4.7.2 Model extensions

The model could be extended to a 3D model with the shape approximated by a slice of a cylinder, in which the nuclei would be represented by spheres, instead of circles, the apical network would be represented by a 2D network of apical points connected by springs and the basement membrane would be represented by a plan. This model would allow to develop a more accurate study of the effect of nuclei packing. Since in 3D there is naturally more free space than in 2D, the current model overestimates the role of packing in the system. A 3D model would also allow to study the effect of a 2D apical network which seems to have a crucial role in the structure and shape of the tissue. In contrast to the model presented in this Chapter, a 2D apical network would be more robust to perturbations caused by the movement of nuclei. We expect that the straightness of the apical network would get closer to the one observed in the real tissue. One could therefore explore the relation between cell behaviour and shape at a deeper level.

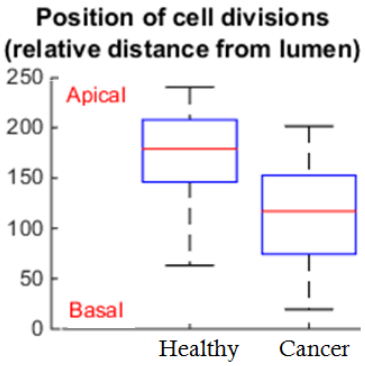
Another extension could be obtained by improving the representation of the nucleus or basement membrane. The shape of the nucleus could be generalized to ellipsoids. In the current model, the basement membrane is represented by a straight line, however, in reality, the membrane consists of a sophisticated matrix constituted by fibers. A model of fibers has been developed in [149] and could be adapted to this case. This modelling extension would allow to study the effect of matrix degradation and cell escape from the tissue, which is associated to cancer metastasis as explained in the next Section.

4.7.3 Studying cancer metastasis and the formation of rosettes

The mechanisms that lead to the first stage of metastasis are similar to the ones occurring during the development of the neural tube in a chick embryo. Therefore the chick embryo has been used as a biological model for studying cancer metastasis [172, 98]. During normal cell division, a cell detaches from the basement membrane, shrinks and divides in two, and then these two new cells elongate and attach again. However, if the cell is cancerous, it has defects, and the elongation and reattachment may not occur. Instead, the cell may keep dividing abnormally and eventually the cancer cells are found attached to each other, forming rosettes. Sometimes a rosette will somehow break through the membrane and start invading the surrounding tissues. The model developed in this Chapter could be used to explore the relation between defects in individual cells and the formation of rosettes. Specifically, two defects that could be incorporated in the model are high proliferation rate and polarity defect, i.e., apical and basal points are not necessarily at opposite sides of the nucleus. The first defect can be included by decreasing the life time of a cell. The second defect can be modelled by decreasing the strength of the alignment force. The model could then be used to explore what could be the contribution of each defect to the formation of rosettes. Preliminary results suggest that a decrease in the life time of the cells prevents the nuclei from reaching the apical domain before division,



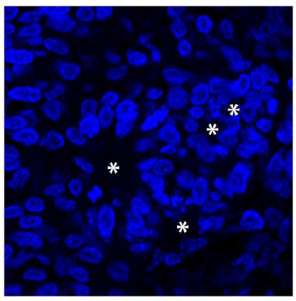
(a) Lab experiment in which Ets1 oncogene is overexpressed to promote tumor formation. Among others, Ets1 leads to increased proliferation.



(b) *In silico* experiment in which cell life time is 16h and 10h in the healthy and cancer tissue, respectively.

Figure 4.11: Comparison of box plots for the height of divisions in the case of a normal tissue and a cancer tissue with a higher rate of proliferation obtained from lab (a) and *in silico* (b) experiments. In both graphs, we observe that the divisions occur near the apical side in the case of the healthy tissue, while in the case of the cancer tissue they occur at any position.

as shown in Figure 4.11b). The same phenomenon is observed in lab experiments, see Figure 4.11a). Another preliminary result on polarity defects, suggests that this defect is enough to generate rosette-type structures, as represented in Figure 4.12.



(a) Lab experiment.

(b) *In silico* experiment.

Figure 4.12: Cluster of cells organized in rosettes obtained *in vivo* (left), by Theveneau's lab, and *in silico* (right). In both images the apical and basal points are situated at the center of the configuration and surround by nuclei.

Part III

Continuum models for congested systems

5

A new continuum theory for incompressible swelling materials

Swelling media (e.g. gels, tumors) are usually described by mechanical constitutive laws (e.g. Hooke or Darcy laws). However, constitutive relations of real swelling media are not well-known. Here, we take an opposite route and consider a simple packing heuristics, i.e. the particles cannot overlap. This approach is based on the same heuristic rules used in Part II in the context of particle based models. We deduce a formula for the equilibrium density under a confining potential. We then consider its evolution when the average particle volume and confining potential depend on time under two additional heuristics: (i) any two particles cannot swap their position; (ii) motion should obey some energy minimization principle. These heuristics determine the medium velocity consistently with the continuity equation. In the direction normal to the potential level sets the velocity is related with that of the level sets while in the parallel direction, it is determined by a Laplace-Beltrami operator on these sets. This complex geometrical feature cannot be recovered using a simple Darcy law.

This work is based on a submitted paper written in collaboration with Pierre Degond[*]*, Sara Merino-Aceituno*[†] *and Mickaël Nahon*[‡]*. My own contributions include the resolution of the minimization problem presented in Section 5.3, the formula for the speed of the moving boundary in Section 5.4, which is crucial for the characterization of the normal component of the velocity of the boundary, and participation in discussions regarding the material presented in the remaining Sections. The determination of the tangential component of the velocity in Section 5.5.3 has been contributed by the other authors and it has been included in this thesis for completeness.*

[*]Department of Mathematics, Imperial College London, UK
[†]School of Mathematical and Physical Sciences, University of Sussex, UK
[‡]École Normale Supérieure de Lyon, France

5.1 Introduction

Swelling or drying media are encountered in many contexts such as chemistry or material science (swelling gels), biology (cancer tumors or growing tissues), geosciences (drying of wetting soil), cooking (dough being cooked), etc. The modelling of swelling or drying media from first principles is difficult due to the complex nature of the materials (cells, mixtures, polymers, etc). Often, they have intermediate properties between solids and liquids or can have genuinely new properties (biological tissues). Modelling of swelling or drying material is very important in view of potential applications in health (tumour growth or tissue development) and other sciences.

Mathematical models have been proposed in the context of tumor growth. Many of these models use a fluid-dynamic approach, and specifically, Darcy's law or some elaboration of it [22, 31, 40, 87, 37]. Mathematically, Darcy's law is expressed by $v = -k\,\nabla p$, where v is the fluid velocity, p is the hydrostatic pressure, ∇ is the spatial gradient and k is a constant named 'hydraulic conductivity'. Darcy's law is derived from Navier-Stokes equation for a fluid subjected to strong friction such as flowing inside a porous medium. However, the use of Darcy's law is not obvious. The article [12] is entirely devoted to the problem of determining the velocity in the mass balance equations (referred to as the "closure problem") and to a phenomenological justification of the use of Darcy's law in tumour growth.

Due to its importance in the clinic, one of the major questions explored in tumor growth modelling is the description of the tumor boundary and how it evolves in time. It naturally leads to the study of free boundary problems [71]. Related to these, the analogy between tumor growth and the free-boundary problem of solidification (the so-called Hele-Shaw problem) has been developed in [88, 146, 145, 147]. In these last series of works, the tumor consists on the region of space where cells have reached the packing density. The tumor is therefore an incompressible medium separated from the outer medium by a moving free boundary which can be calculated through the resolution of an elliptic problem for the pressure in the moving domain of the tumor.

All the previous studies rely on a continuum description of the tumor. However, at the microscopic level, a tumor is made of discrete entities, the cells and various types of "individual-based" microscopic models of tumor growth, where cells are described as discrete entities, have been developed: see in particular [54]. We refer to [158] for a review of the various modelling approaches and to [30] for a comparison of their merits. The connection of the microscopic approach to the macroscopic one through coarse-graining is investigated in [130].

In the present work, we revisit the closure problem and investigate what motion results from the combination of volume-exclusion (or non-overlapping) and growth. In relation to this, we question the validity of Darcy's law once more. Our approach, rather than relying on constitutive relations like hyper-elasticity or Darcy's law, hypothesizes simple heuristic

rules, more likely to be obeyed in generic situations. Here, the main heuristic rule is that particles cannot overlap. In other words, we directly place ourselves in a context akin to the Hele-Shaw limit as developed in [146] and related works cited above. However, as we will see, our conclusions will be different. We also point out that similar heuristic rules have been applied to other domains, such as crowd modelling (see in particular [123]).

We consider a system made of finite-sized particles at equilibrium in a confining external potential constrained by the non-overlapping condition. We refer to [112] for a discussion of the biological relevance of this description and to Chapter 4 for an example of a discrete model developed within this particular setting. We then let the particle volume and the confinement potential vary with time. The particles follow this evolution adiabatically by remaining at any time at a mechanical equilibrium. The question we want to address is what particle motion results from this situation.

Answering this question in full generality at the discrete level is probably out of reach. So, we formulate a similar problem at the continuum level. We assume a continuum density for a population of particles having finite average volume. The particles are confined by an external potential and we assume the particles at mechanical equilibrium. Our first result is to characterize the resulting equilibrium density. Like in the Hele-Shaw type models referred above, the particles occupy a domain of finite extension in space, limited by a level set of the potential. Inside this domain, the density is equal to the maximal (packing) density allowed by their finite size. Outside this domain the density is zero.

Then, we let the system evolve and we compute the continuum velocity. We first determine the component of the velocity that is normal to the potential level sets by using a non-swapping heuristics rule, i.e., particles cannot swap their positions, which is a reasonable assumption in a packed system. To determine the component of the velocity tangent to the potential level sets, we invoke a second heuristics, namely, the particles movement is preferable continuous rather than with jumps which would generate large velocities. In continuum language, this means that the velocity should obey an energy minimization principle. We show that this principle determines the parallel velocity in a unique way as the parallel gradient along the potential level sets of a velocity potential (not to be confused with the confinement potential). This velocity potential is found by inverting a Laplace-Beltrami operator on each of the level sets.

We will show that in general, it is not possible to neglect the tangential component of the velocity. This means that the velocity at the boundary of the medium is not normal to the boundary. By contrast, the Hele-Shaw limit of the tumor models of [146] leads to a velocity at the boundary which is normal to that boundary. Our model provides a different conclusion and consequently, brings new elements in the debate about the validity of the Darcy law, at least in its simple form when the hydraulic conductivity is a scalar.

The medium under consideration bears analogy with a granular material. There has

been considerable literature on granular media and we refer the reader to [15] for a review. Continuum approaches for granular media are mostly based on thermodynamical considerations (see e.g. the seminal work [74]). These approaches rely on the assumption that the system is at equilibrium. However, in complex media such as gels or tumors, there are momentum exchanges with the environment and energy exchanges through (bio)-chemical processes. Since these are extremely difficult to model on a first physical principle basis, we favor a heuristic approach based on the rules as described above.

The Chapter is structured as follows. In Section 5.2 we summarize the main results of our work and provide a detailed discussion and directions for future work. The following sections are devoted to the proofs. The case of the mechanical equilibrium is dealt with in Section 5.3. Then, the time dependent problem is investigated with first the determination of the normal velocity in Section 5.4 and then that of the tangential velocity in Section 5.5. A short conclusion is drawn in Section 5.6.

5.2 Framework, main results and discussion

5.2.1 Motivation: microscopic background

In this section, we motivate our approach by proposing a model of an incompressible swelling medium at the particle level in the same spirit as the models described in Part I and II. We consider a system consisting of N incompressible spherical particles of positions $x_i \in \mathbb{R}^d$, $d \geq 1$, and radii $R_i > 0$, for $i = 1, \dots, N$. The radii are known but the positions are the solutions of a minimization problem. Specifically, we consider that each particle is subject to a potential energy $V(x_i, R_i)$ for a given known energy function $V(x, R)$. For simplicity, we denote by $\mathcal{X} = (x_1, \dots, x_N)$ and $\mathcal{R} = (R_1, \dots, R_N)$. The total energy of the system is the function

$$E_{\mathcal{R}}(\mathcal{X}) = \sum_{i=1}^{N} V(x_i, R_i). \tag{5.1}$$

The first problem we are interested in consists of minimizing the energy (5.1) over a set of admissible configurations $\mathcal{X}$ corresponding to non-overlapping spheres. Specifically, we define the admissible set by

$$\mathcal{A}_{\mathcal{R}} = \big\{ \mathcal{X} \in (\mathbb{R}^d)^N \ \mid \ |x_i - x_j| \geq R_i + R_j, \, \forall i, j \in \{1, \dots, N\}, \, i \neq j \big\}. \tag{5.2}$$

The minimization problem consists of finding $\mathcal{X} \in (\mathbb{R}^d)^N$ which realizes

$$\min_{\mathcal{X} \in \mathcal{A}_{\mathcal{R}}} E_{\mathcal{R}}(\mathcal{X}). \tag{5.3}$$

This pictures the equilibrium configuration of a granular medium made of frictionless spheres in an external potential. Introducing friction or cohesion between the grains is

discarded here and will be investigated in future works. Problem (5.3) has been considered numerically and analytically in Part I. This is a non-convex problem with multiple solutions. We would like to characterize the properties of a generic solution and to this end, we will consider a continuum version of it.

The second problem we consider is the introduction of time evolution dynamics in the system following the same lines as the models developed in Part II. This dynamics is generated by the changes over time of the particles radii $R_i(t)$, which can increase (case of a swelling material) or decrease (case of a drying material). We also allow the potential energy V to depend on time. Here we will suppose that both evolutions are given. Since, the vector of the particle radii $\mathcal{R}(t)$ changes over time, the admissible set $\mathcal{A}_{\mathcal{R}(t)}$ and the potential $V(x,t,R)$ depend on time. Consequently, solutions of (5.3) will also depend on time. Indeed, we assume that the particles stay adiabatically at a minimum of the energy (5.1) and that we can extract a smooth (at least differentiable) trajectory $\mathcal{X}(t)$ among the possible solutions, at least for a small interval of time. The problem is then to find the particle velocities $v_i(t) = \frac{dx_i}{dt}$, or in other words, the vector

$$\mathcal{V}(t) = (v_1(t), \dots, v_N(t)) = \frac{d\mathcal{X}}{dt}(t). \tag{5.4}$$

Again, we discard any friction or cohesion forces between the grains which could alter the time dynamics.

A similar problem has been investigated numerically in [122]. In particular, one possible algorithm is to introduce a time discretization $t^k = k\,\Delta t$ with a time step $\Delta t > 0$ and assume that $\mathcal{X}^k$ is a solution of (5.3) associated to radii $\mathcal{R}^k = \mathcal{R}(t^k)$ and potential function $V^k(x,R) = V(x,t^k,R)$. Then, time is incremented by Δt and a new minimization problem is considered associated to radii $\mathcal{R}^{k+1}$ and potential function V^{k+1}. Obviously, $\mathcal{X}^k$ is not a solution of this new minimization problem. So, a new solution $\mathcal{X}^{k+1}$ is sought. To single out a unique solution among the many possible solutions of the minimization problem, we select the solution $\mathcal{X}^{k+1}$ which has the smallest distance to $\mathcal{X}^k$. In this way, a discrete configuration $\mathcal{X}^{k+1}$ is found, from which a set of discrete velocities

$$\mathcal{V}^k = \frac{\mathcal{X}^{k+1} - \mathcal{X}^k}{\Delta t}, \tag{5.5}$$

is found. The selection principle above leads to the velocity $\mathcal{V}^k$ of smallest possible norm among the possible candidates. The question is whether we can find a simple expression to determine $\mathcal{V}^k$.

Finding a simple answer to this question seems unlikely in the discrete setting, but the problem may be easier to study at the level of a coarse-grained continuum model. So, the goal of this work is to propose such a continuum model and to show that indeed, it is possible to determine these velocities in a unique way. We would like to stress here that it is not our goal to justify the coarse-graining procedure. Rather, we are going to postulate

the problem at the continuum level as an analogue of the problem at the discrete level. The investigation of the passage from the discrete to the continuum problem will be the subject of future work (see also [130] for the coarse-graining of a related model).

5.2.2 General assumptions

We assume a medium made of discrete entities each having finite volume and minimizing a confinement energy subject to a non-overlapping (incompressibility) constraint such as described in Section 5.2.1. Since we are aiming at a continuum description, we do not describe each particle individually but consider their number density $n(x,t)$ and their average volume $\tau(x,t) > 0$, where $x \in \mathbb{R}^d$ is the position in a d-dimensional space (in practice $d = 1$, 2 or 3) and $t \geq 0$ is the time. The non-overlapping constraint (which, at the discrete level, was expressed by the fact that $\mathcal{X}$ must belong to the admissible set $\mathcal{A}_{\mathcal{R}}$) is now expressed by the fact that at any given point in space and time, the volume fraction occupied by the particles $n(x,t)\tau(x,t)$ cannot exceed 1, i.e.

$$n(x,t)\,\tau(x,t) \leq 1. \tag{5.6}$$

Thus, $\tau^{-1}(x,t)$ is the maximal allowed (packing) density of the particles. We assume that $\tau(x,t)$ is a given function of space and time (exactly like in the discrete setting $\mathcal{R}$ was assumed to be a function of time) and that it is defined, positive and finite irrespective of the presence of particles at (x,t). The precise value of $\tau(x,t)$ in practice depends on the modelling context and will be made precise in future work. We also impose that the particle density is nonnegative:

$$n(x,t) \geq 0. \tag{5.7}$$

Additionally, like in the discrete case, we assume that the total number of particles N is fixed, given and is constant in time, i.e.

$$\int_{\mathbb{R}^d} n(x,t)\,dx = N. \tag{5.8}$$

Again, in future work, this assumption will be removed and replaced by a model for the growth or shrinkage of the population.

5.2.3 Mechanical equilibrium

We are first interested by the mechanical equilibrium. Freezing the time variable t for the moment, we assume that there exists a mechanical energy

$$F_t[n] = \int_{\mathbb{R}^d} V(x,t,\tau(x,t))\,n(x,t)dx, \tag{5.9}$$

associated with a given potential $V(x,t,\tau)$, which the particles try to minimize while satisfying the non-overlapping constraint (5.6), the nonnegativity constraint (5.7) and the total mass constraint (5.8). In other words, our goal is to solve the following minimization problem at any given time t:

$$\begin{aligned}&\text{Find } n(\cdot,t):\ x\in\mathbb{R}^d\mapsto n(x,t)\in\mathbb{R}\ \text{ a solution of:}\\&\min\big\{F_t[n(\cdot,t)]\quad|\quad n(\cdot,t)\geq 0,\quad n(\cdot,t)\tau\leq 1\ \text{ and }\ \int_{\mathbb{R}^d} n(x,t)dx=N\big\},\end{aligned}\tag{5.10}$$

for $\tau:(x,t)\in\mathbb{R}^d\times[0,\infty)\mapsto\tau(x,t)\in\mathbb{R}_+$ and $N>0$ given. The potential $V(x,t,\tau)$ is the continuum analog of the discrete potential V of Section 5.2.1 and Eq. (5.9) is nothing but an approximation of Eq. (5.1) when N is large, assuming that the particle positions x_i are drawn randomly, independently and identically according to the probability $N^{-1}\,n(x,t)\,dx$. Obviously, whether this independence assumption holds needs to be proved but we will leave justifications of this question to future work.

We assume that $V\geq 0$. For the simplicity of notations, we define an "effective potential" $W(x,t)$ by

$$W(x,t)=V(x,t,\tau(x,t)).\tag{5.11}$$

We assume that, for all $t\geq 0$, we have

$$W(x,t)\to+\infty\quad\text{ as }\quad|x|\to+\infty.\tag{5.12}$$

In Section 5.3, we will show that, under appropriate conditions on the potential V including (5.12), the solution $n_N(x,t)$ of the minimization problem (5.10) (indexed by the number N of particles in the system) is given by

$$n_N(x,t)=\begin{cases}\dfrac{1}{\tau(x,t)}, & \text{if}\quad x\in\Omega_N(t),\\ 0, & \text{if}\quad x\notin\Omega_N(t),\end{cases}\tag{5.13}$$

where the domain $\Omega_N(t)$ is given by

$$\Omega_N(t)=\{x\in\mathbb{R}^d\mid 0\leq W(x,t)\leq U_N(t)\},\tag{5.14}$$

and $U_N(t)$ is the unique solution of the equation

$$P(U_N(t),t)=N,\tag{5.15}$$

with $P:(u,t)\in[0,\infty)^2\mapsto P(u,t)\in[0,\infty)$ given by

$$P(u,t)=\int_{\{x\in\mathbb{R}^d,\,0\leq W(x,t)\leq u\}}\tau^{-1}(x,t)\,dx.\tag{5.16}$$

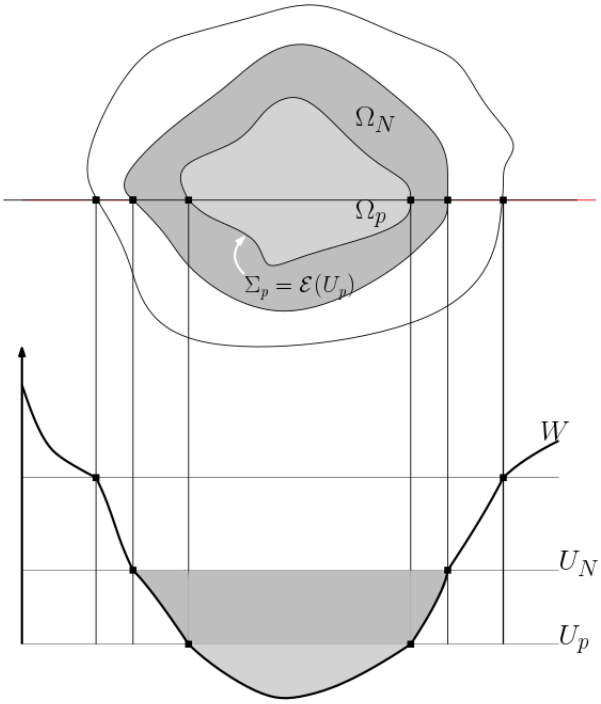


Figure 5.1: Schematics of the filling of the potential level sets. The level set $U_N(t)$ corresponds to the filling of the potential level sets by the entire population of particles N.

Eq. (5.13) shows, that within its support, the density saturates the congestion constraint (5.6), i.e. the density is everywhere equal to the maximal allowed (packing) density $\tau^{-1}(x,t)$. Microscopically, the particles fill all the available space and it is not possible for them to increase the density any further. This is the so-called "packing" or "incompressible" state. To interpret the construction of $\Omega_N(t)$ (formulas (5.14) to (5.16)), we introduce the level sets of the effective potential W. For a given $u \in [0,\infty)$ and time $t \in [0,\infty)$, the level set of $W(\cdot,t)$ corresponding to the value u is defined by:

$$\mathcal{E}_t(u) = \{x \in \mathbb{R}^d,\ W(x,t) = u\}. \tag{5.17}$$

Eq. (5.14) states that $\Omega_N(t)$ is bounded by the level set $\mathcal{E}_t(U_N(t))$. Formula (5.16) defines $P(u,t)$ as the number of particles in the volume limited by the level u. Eq. (5.15) simply states that the level $U_N(t)$ which bounds the domain $\Omega_N(t)$ encloses the total number of particles N, see Fig. 5.1.

Let $n_p(x,t)$ be the solution associated with a different total number of particles $p \geq 0$ with associated support $\Omega_p(t)$. Since P is stricly increasing with respect to u, we have $p < N \Rightarrow U_p(t) < U_N(t)$ and so, with (5.13):

$$p < N \Rightarrow \Omega_p(t) \subsetneq \Omega_N(t) \quad \text{and} \quad n_N(\cdot,t)|_{\Omega_p(t)} = n_p(\cdot,t). \tag{5.18}$$

Additionally, We introduce the domain boundary $\Sigma_p(t)$ of $\Omega_p(t)$. With (5.14) and (5.17), we have

$$\Sigma_p(t) = \partial\Omega_p(t) = \{x \in \mathbb{R}^d \mid W(x,t) = U_p(t)\} = \mathcal{E}_t(U_p(t)). \tag{5.19}$$

This surface will play a crucial role in the definition of the dynamics below. Here, we just remark that, as a consequence of (5.14),

$$\Omega_N(t) = \bigcup_{p\uparrow N} \Sigma_p(t), \tag{5.20}$$

see Fig. 5.1.

5.2.4 Motion under volume growth in non-swapping condition

Now, we turn our attention towards a dynamic situation where the average volume occupied by the particles $\tau(x,t)$ at point (x,t) may vary in time due to either their swelling or drying, described respectively by a time-increasing or decreasing average volume $\tau(x,t)$. We also allow for a possible time-dependence of the confinement potential function $V(x,t,\tau)$. We assume that at any given time t, the medium is at mechanical equilibrium as described in the previous section. So, the time variations of τ and V induce an evolution of the density n and of the material interface $\Omega_N(t)$ in an adiabatic way, i.e. the system follows a trajectory which is a time-continuous sequence of mechanical equilibria, see Fig. 5.2. We are interested by the motion of the material-vacuum interface $\Omega_N(t)$ but also, more importantly, by the motion of the medium itself. More precisely, we would like to define a continuum velocity $v(x,t)$, $x \in \Omega_N(t)$ such that the continuity equation

$$\partial_t n + \nabla \cdot (nv) = 0, \tag{5.21}$$

is satisfied with the solution $n = n_N$ in the domain $\Omega_N(t)$, where ∇ indicates the spatial gradient. Since within $\Omega_N(t)$, $n_N(x,t) = \tau^{-1}(x,t)$ by virtue of (5.13), Eq. (5.21) is an equation for $v(x,t)$, namely:

$$\nabla \cdot (\tau^{-1}(x,t)\, v(x,t)) = -\partial_t \tau^{-1}(x,t), \quad x \in \Omega_N(t), \quad t \geq 0. \tag{5.22}$$

However, it is a scalar equation for the vector quantity $v(x,t)$ and only fully determines v in dimension 1. This is exactly the statement of the "closure problem" discussed in [12]. Here our goal is to determine the velocity $v(x,t)$ fully in any dimension, by following two principles inspired by the microscopic picture, namely, (i) the non-swapping condition and (ii) the principle of smallest displacements. Principle (i) will determine the component of v normal to the family of surfaces $(\Sigma_p(t))_{p\in(0,N]}$ while Principle (ii) will determine its tangential component to these surfaces. We will investigate the consequences of Principle (i) in the present section and defer the use of Principle (ii) to the next section.

The non-swapping principle (Principle (i)) postulates that the level sets of the potential constrain the dynamics of the particles. More precisely, it postulates that two neighboring particles that are on a same level set at one time will continue to be on the same level

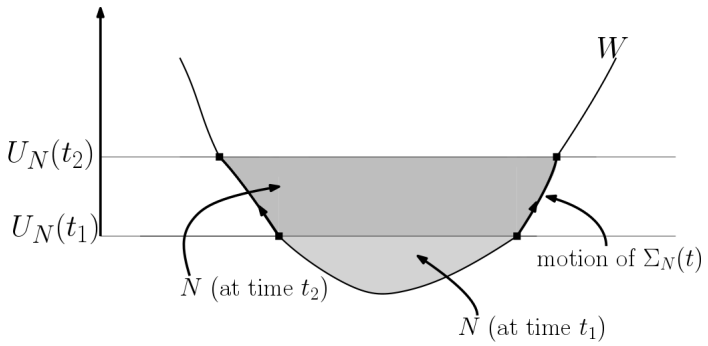


Figure 5.2: Schematics of the motion of the medium between two instants t_1, t_2 where $\tau(\cdot, \tau_2) > \tau(\cdot, \tau_1)$.

set at future times, while those on different level sets will continue to be on different level sets. This non-swapping assumption is a logical consequence of the fact that particles are at a packing state and cannot find enough free space to undertake a swapping maneuver in the normal direction. In dimension $d = 1$, we show that this assumption is always satisfied (given the assumptions made on the data) and consequently, the dynamics is fully determined by the continuity equation. By contrast, in dimension $d \geq 2$, this assumption leads to a non-trivial condition that allows for the unique determination of the component of v normal to the boundary $\Sigma_p(t)$ of $\Omega_p(t)$, for all $p \leq N$. To do so, we introduce

$$\pi(x,t) = P(\, W(x,t)\,, t), \tag{5.23}$$

where W and P are given in (5.11) and in (5.16), respectively. This function gives the number of particles in the volume enclosed by the level set of the effective potential associated with its value at point (x,t), i.e., $u = W(x,t)$ in the definition of P (5.16). By (5.19), we have

$$\Sigma_p(t) = \{x \in \mathbb{R}^d \mid \pi(x,t) = p\} = \pi(\cdot,t)^{-1}(\{p\}), \tag{5.24}$$

so that the family $(\Sigma_p(t))_{0 \leq p \leq N}$ is nothing but the family of level sets of the function $\pi(\cdot,t)$. We assume a non-degeneracy condition: $\nabla\pi(x,t) \neq 0$, for all $(x,t) \in \mathbb{R}^d \times [0,\infty)$. In geometrical language, $\pi(\cdot,t)$ endows $\Omega_N(t)$ with a fiber bundle structure with base space $(0,N]$. The vector

$$\nu(x,t) = \frac{\nabla\pi(x,t)}{|\nabla\pi(x,t)|}, \tag{5.25}$$

defines the outward unit normal to $\Sigma_p(t)$ at x with $p = \pi(x,t)$. We can decompose the velocity vector v as follows:

$$v(x,t) = v_\perp(x,t) + v_\parallel(x,t), \quad v_\perp(x,t) = \big((v\cdot\nu)\,\nu\big)(x,t), \quad v_\parallel(x,t)\cdot\nu(x,t) = 0, \tag{5.26}$$

for all $x \in \Omega_N(t)$, $t \in [0,\infty)$. In the sequel, $v_\perp = |v_\perp|\nu$ will be referred to as the normal velocity (with respect to the surface Σ_p with $p = \pi(x,t)$) and $v_\parallel$ as the tangential velocity.

The main consequence of the non-swapping assumption is that in dimension $d \geq 2$, it

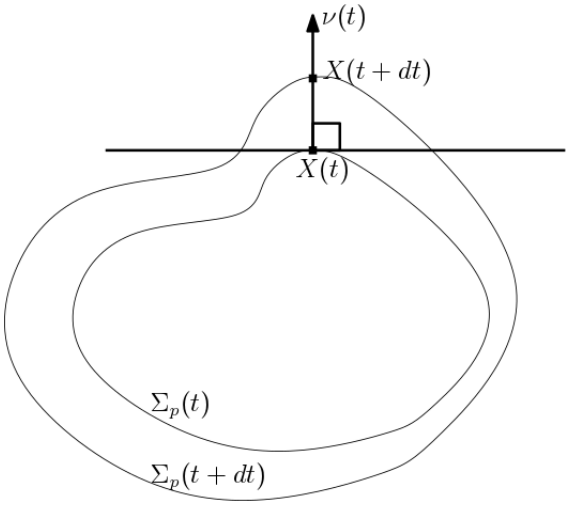


Figure 5.3: Medium velocity in the normal direction is the velocity of Σ_p, i.e. $w_\perp = \frac{1}{dt}\left(X(t+dt) - X(t)\right) \cdot \nu(t)$ (see Definition 39).

leads to the full determination of the modulus of the normal velocity $|v_\perp| = w_\perp$ as follows:

$$w_\perp(x,t) = -\frac{\partial_t \pi(x,t)}{|\nabla \pi(x,t)|}, \quad x \in \Omega_N(t), \quad t \geq 0. \tag{5.27}$$

This is nothing but the velocity of $\Sigma_p(t)$ in the normal direction. The interpretation is that, due to the non-swapping assumption, any particle located in the infinitesimal layer between $\Sigma_p(t)$ and $\Sigma_{p+\delta p}(t)$ with $\delta p \ll 1$ must remain in this layer and therefore, has to move with the velocity of $\Sigma_p(t)$, see Fig. 5.3.

In Section 5.4, we prove that, for any velocity field satisfying (5.27), the left-hand side of the continuity equation (5.21) averaged on $\Sigma_p(t)$ is identically zero for any $p \leq N$ and any $t \geq 0$, namely

$$\left\langle \, \delta \circ \big(\pi(\cdot,t) - p\big) \, , \, \big(\partial_t n + \nabla \cdot (nv)\big)(\cdot,t) \, \right\rangle = 0, \tag{5.28}$$

where $\langle \cdot, \cdot \rangle$ is the duality bracket between a distribution and a smooth function. To interpret the Dirac delta in the expression above, we recall the following formula, a consequence of the so-called coarea formula:

$$\left\langle \, \delta \circ \psi \, , \, f \, \right\rangle = \int_{\{\psi(x)=0\}} f(x) \frac{dS(x)}{|\nabla \psi(x)|}, \tag{5.29}$$

for any smooth functions $x \in \mathbb{R}^d \mapsto f(x)$, $\psi(x) \in \mathbb{R}$, where $dS(x)$ is the Euclidean surface element on the level set $\{x \in \mathbb{R}^d,\, \psi(x) = 0\}$. The notation $(\cdot, t)$ is there to remind that the time variable t is fixed when evaluating the duality bracket in (5.28). Eq. (5.28) will be an important condition for determining the tangential velocity $v_\parallel$ in the next section.

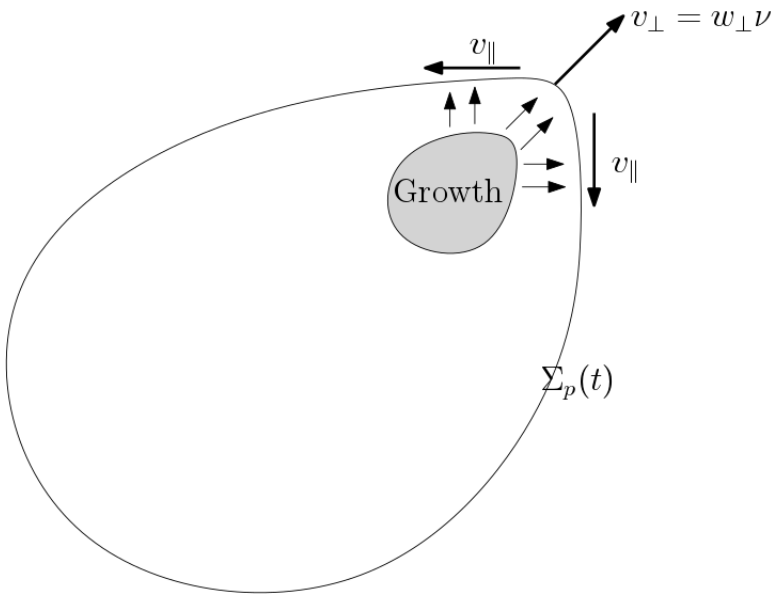


Figure 5.4: Illustration of a need for a non-zero tangential velocity $v_\parallel$.

5.2.5 Tangential velocity

To determine the tangential velocity $v_\parallel$, we apply the principle of smallest displacements (Principle (ii), see previous section). This principle suggests to determine the velocity $v_\parallel$ as the solution of a convenient energy minimization principle. It is the continuum counterpart of the principle set at the microscopic level in Section 5.2.1, which suggested to look for the smallest velocity $\mathcal{V}^k$ among the possible ones. In the present section, we summarize the conclusions of this approach ; details and proofs can be found in Section 5.5.

First, let us make a special mention of dimension 1, as in this case, there is no tangential velocity. So, a natural question is whether Eq. (5.27) is compatible with the continuity equation (5.21). In Section 5.5.1, we will show that this is indeed the case. This will be a consequence of (5.28).

Second, we point out that in dimension $d \geq 2$, we do need a non-zero tangential velocity $v_\parallel$ in general. Indeed, even if the choice $v = v_\perp \nu$ with $v_\perp$ as in (5.27) satisfies (5.28), it does not necessarily satisfy the continuity equation (5.21). In Section 5.5.2, we will give a two-dimensional counter-example where this is indeed not true, see Fig. 5.4.

So, if $d \geq 2$, in order to satisfy the continuity equation, the velocity has to incorporate a non-trivial parallel component $v_\parallel$. Using (5.26), the continuity equation (5.22) can be written

$$\nabla \cdot (\tau^{-1} v_\parallel) = f, \qquad f := -\partial_t \tau^{-1} - \nabla \cdot (\tau^{-1} w_\perp \nu), \qquad x \in \Omega_N(t),\ t \geq 0, \tag{5.30}$$

and appears as a constraint on $v_\parallel$. Eq. (5.28) tells us that, for each $p \leq N$ and $t \geq 0$, the average of f on the level set $\Sigma_p(t)$ (defined in (5.24)) is zero, namely,

$$\left\langle \delta \circ \big(\pi(\cdot,t) - p\big)\,,\, f(\cdot,t) \right\rangle = 0, \qquad \forall (p,t) \in (0,N] \times [0,\infty). \tag{5.31}$$

In section 5.5.3, we show that (5.31) is a necessary condition for the existence of a solution to (5.30). It is also a sufficient condition. However, in order to guarantee the uniqueness of the solution, we need to impose an additional constraint.

Here, we add the condition that $v_\parallel$ corresponds to the minimal displacement on each of the level sets $\Sigma_p(t)$. In other words, we search for the vector fields $v_\parallel$ that minimize the parallel kinetic energy

$$\mathcal{K}_{p,t}[v_\parallel] = \left\langle \delta \circ \big(\pi(\cdot,t) - p\big)\,,\, |v_\parallel(\cdot,t)|^2 \right\rangle, \qquad \forall (p,t) \in (0,N] \times [0,\infty), \tag{5.32}$$

on all surfaces $\Sigma_p(t)$, i.e.

$$v_\parallel \in \arg\min\{\, \mathcal{K}_{p,t}[w_\parallel],\ w_\parallel \text{ s.t. } \nabla \cdot (\tau^{-1}\, v_\parallel) = f \,\}, \qquad \forall (p,t) \in (0,N] \times [0,\infty), \tag{5.33}$$

where we denote by arg min the set of minimizers of the expression inside the curly brackets. The expression (5.32) is nothing but the integral of the parallel kinetic energy density on the surface $\Sigma_p(t)$. Indeed, the parallel kinetic energy of a particle of volume τ is proportional to $\tau |v_\parallel|^2$ but the density of such particles is proportional to τ^{-1}. The contributions of the particle volume τ cancel, which leads to the expression (5.32).

In section 5.5.3, we show that such vector fields are necessarily surface gradients on the level set $\Sigma_p(t)$ of scalar functions. Specifically, we will show that (5.33) implies that there exists a scalar function $\theta(x,t)$, such that:

$$v_\parallel(x,t) = -\nabla_\parallel \theta(x,t), \qquad \nabla_\parallel \theta(x,t) := \nabla\theta(x,t) - \big(\nabla\theta(x,t) \cdot \nu(x,t)\big)\, \nu(x,t), \tag{5.34}$$

where $\nabla_\parallel$ is the tangential gradient parallel to the level sets $\Sigma_p(t)$. With this condition, (5.30) becomes an elliptic equation for θ on each level set surface $\Sigma_p(t)$, written as

$$-\nabla_\parallel \cdot (\tau^{-1}\, \nabla_\parallel \theta) = f, \qquad x \in \Omega_N(t), \quad t \geq 0, \tag{5.35}$$

In section 5.5.3, this equation will be shown to have a unique solution in an appropriate function space, provided that (5.31) holds and that θ is sought with average zero on each level surface, namely

$$\left\langle \delta \circ \big(\pi(\cdot,t) - p\big)\,,\, \theta(\cdot,t) \right\rangle = 0, \qquad \forall (p,t) \in (0,N] \times [0,\infty). \tag{5.36}$$

Indeed, (5.35) can be reformulated as the inversion of a Laplace-Beltrami operator on each of the level surfaces $\Sigma_p(t)$. Standard differential geometry (see [72], Section 4.D.2) asserts that if the solution is sought in the subspace $H^1_0(\Sigma_p(t))$ of the Sobolev space $H^1(\Sigma_p(t))$ consisting of functions satisfying the additional constraint (5.36), this inversion has a unique solution.

If the problem has spherical symmetry, i.e. if there exists $\mathcal{V} : (r,t,\tau) \in [0,\infty)^3 \mapsto \mathcal{V}(r,t,\tau) \in [0,\infty)$ and $\mathcal{T} : (r,t) \in [0,\infty)^2 \mapsto \mathcal{T}(r,t) \in [0,\infty)$ such that $V(x,t,\tau) =$

$\mathcal{V}(|x|, t, \tau)$, $\tau(x,t) = \mathcal{T}(|x|, t)$, then the unique solution of (5.35), (5.36) is $\theta = 0$, which shows that in this case $v_{\parallel} = 0$ and $v = v_{\perp}\nu$.

5.2.6 Discussion

First, we discuss the stationary equilibrium exposed at Section 5.2.3. The result given in (5.13), (5.14) proves that the solution of the minimization problem is unique, contrary to the discrete case exposed in Section 5.2.1. These formulas show that the particles gradually fill the energy level sets of the effective potential W by increasing values while keeping the non-overlapping condition saturated (i.e. the density being equal to the packing density). Indeed, the effective potential W tends to bring all particles towards its points of global minimum. However, the non-overlapping constraint prevents the particles to pile up at these points and forces them to occupy increasingly higher potential values. They do so until the total number of particles has been exhausted. When this happens, the medium has reached its outer boundary and is therefore limited by the level set that encloses a number of particles exactly equal to the total number N of available particles in the system (see Eq. (5.15)).

This can be compared to the process by which electrons fill energy levels in a perfect crystal at zero temperature. Electrons fill the crystal energy levels by increasing energy due to Pauli's exclusion principle which prevents a given energy level to be occupied twice. The energy corresponding to the last occupied energy level is called the Fermi energy. The present picture is similar and $U_N(t)$ (Eq. (5.15)) could be viewed as the Fermi energy of our medium. The measure $\frac{dP}{du}(u,t)\,du$ (see Eq. (5.16)) which can be interpreted as the infinitesimal number of particles in a small energy interval du around energy u is similar to what solid-state physicists call the density-of-states, see Fig. 5.1.

We now comment on the time-dependent case and the determination of the velocity in Sections 5.2.4 and 5.2.5. In these two sections, we provide an answer to the "closure problem" [12], i.e. the problem of determining the velocity field consistent with the continuity equation (5.21). This answer is different from the classical one relying on Darcy's law. Consequences of the use of Darcy's law for incompressible swelling materials can be found e.g. in [146]. One of these is that, at the medium boundary, the continuum velocity is normal to the boundary. In the framework presented here, the velocity at the medium boundary does not have to be (and is not in general) normal to the boundary, due to the presence of a non-trivial tangential velocity component. The presence of this tangential component allows the material to move along regions of constant energy. Since the dynamics is smooth, this situation may allow a faster, and perhaps a more radially asymmetric evolution of the boundary of the material when compared with classical models. This discrepancy with Darcy's law may result from confinement by the external potential V acting independently from the growth source modelled by $d\tau/dt$, see Fig. 5.4. In [146], the confinement pressure is directly computed from the growth source term with-

out any external potential V. Situations where confinement arises from external factors may be of importance for instance in tumor modelling when the tumor is confined by the surrounding tissue.

The model presented here is a building block towards a more realistic description of swelling materials such as swelling gels or tumours. This new modelling approach opens many exciting new research directions, from theory to numerics and modelling to applications. A (non-exhaustive) list of future directions which will be investigated in forthcoming works include the following: adding cell division; consider a potential V that involves a contribution from particle interaction such as attachment between nearby cells; coupling with chemical fields; introduction of boundary fuzziness; introduction of a statistical description of particle volume sizes leading to a kinetic equation; taking into account multiple particle species; derivation from a microscopic model by coarse-graining; numerical approximation and applications to practical systems.

The following three sections provide the mathematical foundations of the results exposed so far.

5.3 Equilibrium through confinement subject to volume exclusion constraint

In the present section, we provide the mathematical background to the conclusions exposed in Section 5.2.3, i.e. we determine the equilibrium configuration of the particles at a given time t. Throughout this section, t is only a parameter, and so we will omit it in the expression of all the variables. The equilibrium configuration corresponds to minimizing the confinement energy $F[n]$ given by (5.9) subject to the volume exclusion constraint (5.6), the nonnegativity constraint (5.7) and the total number of particles constraint (5.8). Therefore, we are led to solving the minimization problem (5.10) which we rewrite as follows since we omit the time-dependence:

$$\begin{aligned} &\text{Find } n: \ x \in \mathbb{R}^d \mapsto n(x) \in \mathbb{R} \ \text{ a solution of:} \\ &\min\big\{F[n] \quad | \quad n \geq 0, \quad n\tau \leq 1 \ \text{ and } \ \int_{\mathbb{R}^d} n(x)dx = N\big\}, \end{aligned} \tag{5.37}$$

for $\tau : x \in \mathbb{R}^d \mapsto \tau(x) \in \mathbb{R}_+$ and $N > 0$ given. We recall the expressions (5.11) of the effective potential W and write $W = W(x)$ as we ignore the dependence with respect to t. We also recall the definition (5.17) of the level set of W associated to the level value u and we denote this level set by $\mathcal{E}(u)$, again ignoring the time-dependence. In this section we prove the following:

Theorem 37. *Assume the following:*
(i) the functions $x \in \mathbb{R}^d \mapsto W(x) \in \mathbb{R}$ and $x \in \mathbb{R}^d \mapsto \tau^{-1}(x) \in \mathbb{R}$ are smooth ;

(ii) $W(x) \geq 0$, $\forall x \in \mathbb{R}^d$;
(iii) $0 < \tau(x) < \infty$ *for all* $x \in \mathbb{R}^d$;
(iv) $|\nabla W(x)| < \infty$, *for all* $x \in \mathbb{R}^d$;
(v) for all $u > 0$, *the level sets* $\mathcal{E}(u)$ *are compact and have strictly positive* $d-1$ *Lebesgue surface measure;*
(vi) $x = 0$ *is the only critical point of* W *and* $W(0) = 0$*;*
(vii) $W(x) \to +\infty$ *as* $|x| \to +\infty$;
(viii) $\int_{\mathbb{R}^d} \tau^{-1}(x)\, dx > N$*;*
then, the solution of the minimization problem (5.37) is unique and given by (5.13) with the set Ω *given by (5.14)-(5.16).*

Remark 5.3.1. *(i) That* $\mathcal{E}(u)$ *is compact for all* $u > 0$ *(see Assumption (v)) follows from Assumption (vii). However, that they have strictly positive* $d-1$ *dimensional measure does not follow from Assumption (vii). Conversely, Assumption (vii) does not follow from the compactness of* $\mathcal{E}(u)$*.*

(ii) Differentiating expression (5.23) with respect to x*, we obtain:*

$$\nabla \pi(x,t) = \frac{dP}{du}(W(x,t),t)\, \nabla W(x,t). \tag{5.38}$$

By assumption (vi), $\nabla W(x,t) \neq 0$ *for* $x \neq 0$ *and, as we will see in the proof of Th. 37, Eq. (5.48), it holds that*

$$\frac{dP}{du}(u) > 0, \ \forall u > 0.$$

Therefore, from (5.38) we conclude that

$$\nabla \pi(x,t) \neq 0, \quad \text{for } x \neq 0, \tag{5.39}$$

which is a non-degeneracy condition that we will use in the sequel. Moreover, by Assumption (i), using Eq. (5.23), we have that π *is also smooth.*

(iii) A more general form of the coarea formula (5.41) (see Ref. [62]) would allow us to extend the results with weaker assumptions than (vi) or without having to assume that $\nabla \pi \neq 0$*. However, to keep the presentation simple, we do not follow this path here. Indeed, with assumption (vi) we ensure that* Ω *stays connected. If we had, say, two connected components, the global minimisation problem (5.37) would fix the number of particles in each of the connected components, which is unrealistic, as we may expect that these two numbers could a priori be chosen independently.*

Before proving Theorem 37 we first prove the following:

Lemma 38. *Suppose the assumptions of Theorem 37 hold. Then, a solution* n *of the*

minimization problem (5.37) is such that, for all $x \in \mathbb{R}^d$,

$$\textit{either} \quad n(x)\,\tau(x) = 1 \quad \textit{or} \quad n(x) = 0. \tag{5.40}$$

Proof: We start by recalling the theory for convex optimization exposed in Section 0.3.3.1 in order to prove the existence of Lagrange multipliers associated to a solution of the minimization problem (5.37). First note that F and the functions associated to the constraints are continuous and Gateaux differentiable (see [174] for a proof of differentiability). Moreover, all these functions are linear and, therefore, the problem (5.37) is convex (see Section 0.3.1.2, definition 1). Second, under hypothesis (*viii*) of theorem 37, we have that $N < M$, with $M = \int_{\mathbb{R}^d} \tau^{-1} dx$. Consequently, the point

$$n_1 = \frac{N}{M}\tau^{-1},$$

is a strictly admissible point of (5.37), so the Slater condition holds true.

Let n be a solution to the minimization problem (5.37). Then, recalling theorem 5 from Section 0.3.3.1, there exist three Lagrange-Kuhn-Tucker multipliers λ, μ and ν, where $\mu \in \mathbb{R}$ and $\lambda = \lambda(x) \geq 0$ and $\nu = \nu(x) \geq 0$ are functions satisfying: (i) $\lambda(x) = 0$ for all x such that $n(x)\,\tau(x) < 1$; and (ii) $\nu(x) = 0$ for all x such that $n(x) > 0$; such that the Euler-Lagrange equations hold:

$$\int W(x)\,\delta n(x)\,dx = -\int \lambda(x)\,\tau(x)\,\delta n(x)\,dx + \int \nu(x)\,\delta n(x)\,dx + \mu \int \delta n(x)\,dx,$$

for all small variations $\delta n(x)$ of $n(x)$. The last term corresponds to the constraint on the total mass being equal to N. It follows that

$$W(x) = -\lambda(x)\,\tau(x) + \nu(x) + \mu.$$

Now, suppose that $n(x')\,\tau(x') < 1$ and $n(x') > 0$ for x' in a neighbourhood $\mathcal{U}$ of a point x. Then, $\lambda = 0$ and $\nu = 0$ in $\mathcal{U}$ and

$$W(x) = \mu = \text{Constant}, \quad \forall x \in \mathcal{U}.$$

This occurrence is ruled out by Assumption (vi) of Theorem 37. Therefore, Eq. (5.40) must be verified. Now, thanks to condition (viii) in Th. 37 this is an admissible solution, which ends the proof of the Lemma. ∎

Before turning to the proof of Theorem 37, we recall the coarea formula in its general

form (formula (5.29) is a particular case involving the Dirac delta):

$$\int_{\mathbb{R}^d} f(x)\,dx = \int_{\psi(\mathbb{R}^d)} \Big(\int_{\{\psi(x)=u\}} f(x) \frac{dS_u(x)}{|\nabla\psi(x)|} \Big)\, du, \tag{5.41}$$

where $x \in \mathbb{R}^d \mapsto \psi(x)$, $f(x) \in \mathbb{R}$ are smooth functions and $dS_u(x)$ is the Euclidean surface element on the codimension-1 manifold $\{\psi(x) = u\}$ and $\nabla\psi$ is nowhere zero (these assumptions can be relaxed, see [62]). With (5.29), we can also write (5.41) as

$$\int_{\mathbb{R}^d} f(x)\,dx = \int_{\psi(\mathbb{R}^d)} \big\langle \delta \circ (\psi - u)\,,\, f \big\rangle\, du. \tag{5.42}$$

In particular, we have

$$\int_{\mathbb{R}^d} f(x)\,(g\circ\psi)(x)\,dx = \int_{\psi(\mathbb{R}^d)} \big\langle \delta \circ (\psi - u)\,,\, f \big\rangle\, g(u)\,du, \tag{5.43}$$

where $g : \psi(\mathbb{R}^d) \mapsto \mathbb{R}$ is a smooth function.

Proof of Proposition 37. Thanks to Lemma 38, any solution of (5.37) is of the form (5.13) where the only unknown is the set Ω. We denote by χ_Ω the indicator function of the set Ω (we recall that the indicator function of a set A is the function that takes the value 1 on A and the value 0 on its complement set). Then, by the coarea formula (5.43) applied with $f = \tau^{-1}\chi_\Omega$, $g(u) = u$ and $\psi = W$, we get, since $n(x) = \tau^{-1}(x)$ on Ω:

$$\begin{aligned} F[n] &= \int_\Omega W(x)\,\tau^{-1}(x)\,dx \\ &= \int_{\mathbb{R}^d} W(x)\,\tau^{-1}(x)\,\chi_\Omega(x)\,dx \\ &= \int_0^{+\infty} \big\langle\, \delta\circ(W-u)\,,\, \tau^{-1}\,\chi_\Omega \,\big\rangle\, u\,du. \end{aligned} \tag{5.44}$$

Here the integration with respect to u can be taken over $[0,\infty)$ thanks to Assumption (ii) of Theorem 37. We recall that, following (5.29)

$$\big\langle\, \delta\circ(W-u)\,,\, \tau^{-1}\ \chi_\Omega \,\big\rangle = \int_{\mathcal{E}(u)} \frac{\tau^{-1}(x)\,\chi_\Omega(x)\,dS_u(x)}{|\nabla W(x)|},$$

where $dS_u(x)$ is the Euclidean surface element on $\mathcal{E}(u)$ and $\mathcal{E}(u)$ is defined at (5.17). Consequently, the only values of $\chi_\Omega(x)$ that enter the integral (5.44) for a fixed value of u are those taken on $\mathcal{E}(u)$. We claim that the minimum of $F[n]$ is reached if and only if the following is satisfied: (i) $\chi_\Omega(x)$ (which is equal to 0 or 1) is constant (i.e. either constantly 0 or constantly 1) on any level set $\mathcal{E}(u)$ for all $u \geq 0$; (ii) there exists $U > 0$ such that $\chi_\Omega(x) = 1$ on $\mathcal{E}(u)$ for all u such that $0 \leq u \leq U$ and $\chi_\Omega(x) = 0$ for $u \geq U$.

Equivalently, these two conditions put together mean that $\chi_\Omega(x)$ can be written:

$$\chi_\Omega(x) = \chi_{[0,U]}(W(x)), \quad \text{i.e.} \quad \chi_\Omega = \chi_{[0,U]} \circ W. \tag{5.45}$$

It follows that (thanks to (5.43)

$$\left\langle \delta \circ (W-u)\,,\, \tau^{-1}\,\chi_\Omega \right\rangle = \chi_{[0,U]}(u)\,\left\langle \delta \circ (W-u)\,,\, \tau^{-1} \right\rangle, \tag{5.46}$$

and

$$F[n] = \int_0^U \left\langle \delta \circ (W-u)\,,\, \tau^{-1} \right\rangle\, u\, du. \tag{5.47}$$

Assuming this result for a while, i.e., that U satisfying (5.45) exists, we show that U is uniquely determined by the total number of particles constraint (5.8). Using (5.45) and the fact that on Ω, $n(x) = \tau^{-1}(x)$, we can compute the total mass as follows:

$$\begin{aligned} N &= \int_\Omega \tau^{-1}(x)\, dx \\ &= \int_{\mathbb{R}^d} \tau^{-1}(x)\, \chi_\Omega(x)\, dx \\ &= \int_{\mathbb{R}^d} \tau^{-1}(x)\ \chi_{[0,U]}(W(x))\, dx \\ &= \int_{\{x\in\mathbb{R}^d\,,\ 0\le W(x)\le U\}} \tau^{-1}(x)\, dx \\ &= P(U), \end{aligned}$$

where the function P (for fixed time t) is defined by (5.16). This leads to Eq. (5.15) for the determination of U. Note that $P(U) < \infty$ for any $U \ge 0$ by Assumption (vii).

Now, Eq. (5.15) has a unique solution. Indeed, using the coarea formula again, we have

$$P(u) = \int_0^u \left\langle \delta \circ (W-u')\,,\, \tau^{-1} \right\rangle du'.$$

Therefore, using (5.29) and recalling the definition (5.17) of $\mathcal{E}(u)$, we have

$$\begin{aligned} \frac{dP}{du}(u) &= \left\langle \delta \circ (W-u)\,,\, \tau^{-1} \right\rangle \\ &= \int_{\mathcal{E}(u)} \tau^{-1}(x)\, \frac{dS_u(x)}{|\nabla W(x)|}. \end{aligned}$$

From Assumptions (i) and (iii) to (iv) and (vi) of Theorem (37), there exists $C_u > 0$ such that $\tau^{-1}(x)\,|\nabla W(x)|^{-1} \ge C_u > 0$ on $\mathcal{E}(u)$. Thus, by Assumption (v) of Theorem (37),

$$\frac{dP}{du}(u) \ge C_u \int_{\mathcal{E}(u)} dS_u(x) > 0. \tag{5.48}$$

Consequently, P is a strictly increasing function and there exists a unique $u = U$ such that (5.15) holds.

We now show (5.14). Denote by Ω_0 the set defined by (5.14) and by n_0 the corresponding density given by (5.13). Taking χ_Ω not of the form (5.14), we show that the corresponding density n has energy strictly larger than that of n_0, i.e. $F[n] > F[n_0]$. This incidentally shows the uniqueness of the solution of the minimization problem as from Lemma 38, it must be of the form (5.13) for some set Ω and if $\Omega \neq \Omega_0$, then, its energy is strictly larger than that obtained with Ω_0.

Taking $\Omega \neq \Omega_0$ means that at least one of the subsets

$$\omega_1 = \{x \in \mathbb{R}^d, \text{ such that } W(x) \leq U \text{ and } \chi_\Omega = 0\},$$

or

$$\omega_2 = \{x \in \mathbb{R}^d, \text{ such that } W(x) > U \text{ and } \chi_\Omega = 1\},$$

contains a non-zero number of particles (i.e. has non-zero measure for the measure $\tau^{-1}(x)\,dx$). We now show that they both contain a non-zero number of particles and that these numbers are the same by the total number of particles constraint (5.8). Indeed, we note that

$$\Omega_0 \setminus \omega_1 = \Omega \setminus \omega_2 = \{x \in \mathbb{R}^d, \text{ such that } W(x) \leq U \text{ and } \chi_\Omega = 1\}. \tag{5.49}$$

Denote this set by $\tilde{\omega}$. Then, by the constraint (5.8), we can write:

$$N = \int_{\Omega_0} \tau^{-1}(x)\,dx = \int_{\Omega} \tau^{-1}(x)\,dx.$$

Decomposing the first integral on ω_1 and $\tilde{\omega}$ (which form a partition of Ω_0 by (5.49)) and the second one on ω_2 and $\tilde{\omega}$ (which similarly form a partition of Ω), we get:

$$\int_{\omega_1} \tau^{-1}(x)\,dx + \int_{\tilde{\omega}} \tau^{-1}(x)\,dx = \int_{\omega_2} \tau^{-1}(x)\,dx + \int_{\tilde{\omega}} \tau^{-1}(x)\,dx,$$

and consequently

$$\int_{\omega_1} \tau^{-1}(x)\,dx = \int_{\omega_2} \tau^{-1}(x)\,dx, \tag{5.50}$$

showing that the number of particles contained in ω_1 and ω_2 are the same. Note that, by the coarea formula (5.42), we can re-write (5.50) according to:

$$\int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, (\chi_{\omega_2} - \chi_{\omega_1})\,\tau^{-1} \right\rangle du = 0. \tag{5.51}$$

Now, we have, thanks to (5.44)

$$F[n] - F[n_0] = \int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, (\chi_\Omega - \chi_{\Omega_0})\,\tau^{-1} \right\rangle u\, du. \tag{5.52}$$

We note that

$$\chi_{\Omega_0} = \chi_{\omega_1} + \chi_{\tilde{\omega}}, \quad \chi_\Omega = \chi_{\omega_2} + \chi_{\tilde{\omega}}.$$

So, (5.52) is written

$$F[n] - F[n_0] = \int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, (\chi_{\omega_2} - \chi_{\omega_1})\,\tau^{-1} \right\rangle u\, du. \tag{5.53}$$

But we have

$$\omega_2 \subset \{x \in \mathbb{R}^d,\ W(x) > U\}, \quad \omega_1 \subset \{x \in \mathbb{R}^d,\ W(x) \leq U\}.$$

So, we can write

$$\begin{aligned} \int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_2}\,\tau^{-1} \right\rangle u\, du &= \int_U^{+\infty} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_2}\,\tau^{-1} \right\rangle u\, du \\ &> U \int_U^{+\infty} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_2}\,\tau^{-1} \right\rangle du \\ &= U \int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_2}\,\tau^{-1} \right\rangle du, \end{aligned} \tag{5.54}$$

and similarly

$$\begin{aligned} \int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_1}\,\tau^{-1} \right\rangle u\, du &= \int_0^{U} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_1}\,\tau^{-1} \right\rangle u\, du \\ &\leq U \int_0^{U} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_1}\,\tau^{-1} \right\rangle du \\ &= U \int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, \chi_{\omega_1}\,\tau^{-1} \right\rangle du, \end{aligned} \tag{5.55}$$

Therefore,

$$F[n] - F[n_0] > U \int_0^{+\infty} \left\langle \delta \circ (W - u)\,,\, (\chi_{\omega_2} - \chi_{\omega_1})\,\tau^{-1} \right\rangle du. \tag{5.56}$$

But the integral at the right-hand side of (5.56) is equal to zero by (5.51). Consequently, we get

$$F[n] > F[n_0],$$

which is the result to be proved. Note that the proof relies on the fact that the inequality

in (5.54) is strict. This is only true if the support of the function

$$u \mapsto \left\langle \, \delta \circ (W - u) \, , \, \chi_{\omega_2} \tau^{-1} \, \right\rangle,$$

is not reduced to $\{U\}$. But if this is the case, since the involved function is smooth, this means that it is identically equal to zero. This implies that

$$\int_0^\infty \left\langle \, \delta \circ (W - u) \, , \, \chi_{\omega_2} \, \tau^{-1} \, \right\rangle du = 0,$$

and this is the total number of particles in ω_2. But if there are no particles contained in ω_2, that means that all particles are contained in Ω_0 and therefore $\Omega = \Omega_0$. So, as soon as $\Omega \neq \Omega_0$, we have a strict inequality in (5.54). This ends the proof of Prop. 37. ∎

Remark 5.3.2. *The interpretation of (5.47) is as follows. The measure*

$$dN(u) := \frac{dP}{du}(u)\, du = \left\langle \, \delta \circ (W - u) \, , \, \tau^{-1} \, \right\rangle du,$$

is the number of particles comprised between the level sets $\mathcal{E}(u)$ and $\mathcal{E}(u + du)$ (similar to the density-of-states in solid-state physics, see Section 5.2. In this layer, the effective potential has value u. So, (5.47) expresses that we get the total energy by summing the values of the effective potential u associated to the level set $\mathcal{E}(u)$ between 0 and U, weighted by the number density of particles in this level set.

5.4 Continuum velocity under non-swapping constraint

In this section, we turn our attention to a time-dynamic situation, and provide the mathematical framework to the results described in Section 5.2.4. We consider that the average volume τ of the underlying particles in the continuum medium as well as the potential function V may evolve in time. However, we assume that, during this evolution, the medium stays at mechanical equilibrium under the antagonistic influences of congestion and the volume exclusion constraint at any time. Due to the time-variation of τ and V the particle density $n(x, t)$ will change and we are interested in finding the velocity field $v(x, t)$ of this continuum medium. Such velocity must satisfy the continuity equation (5.21). However, this equation is a scalar equation and can only determine the vector quantity v in dimension one. In dimension more than 2, we need additional physical assumptions to determine v. Here, we examine what additional information on v we can get from assuming that the underlying particles cannot swap their positions. We refer to Section 5.2.4 for a justification of the non-swapping assumption.

In this section, by contrast to the previous one, we restore the time-dependence of all

the quantities involved in the minimization of the mechanical energy (5.9) subject to the constraints (5.6), (5.7), (5.8). We recall that under the assumptions of Theorem 37, the particle density $n_N(x,t)$, the unique solution of this constrained minimization problem, is given by (5.13), where the domain $\Omega_N(t)$ is given by (5.14)-(5.16). We also recall that in dimension $d \geq 2$, the constraint that the particles cannot swap their positions implies that those contained in the layer between two neighbouring level sets $\Sigma_p(t)$ and $\Sigma_{p+\delta p}(t)$ with $\delta p \ll 1$ at time t will remain in this layer at all times. Such particles must move with the layer, i.e. their normal velocity to the layer must be that of the layer or, in other words, that of the boundary $\Sigma_p(t)$.

To express this velocity, we recall the expression (5.23) of the function $\pi(x,t)$ such that $x \in \Sigma_{\pi(x,t)}(t)$. The function $\pi(x,t)$ is the number of particles in the volume enclosed by the level set of the effective potential W associated with the level value $W(x,t)$. By Eq. (5.24) we also have that $\Sigma_p(t)$ is the level set of the function $\pi(\cdot,t)$. We assume the non-degeneracy condition:

$$\nabla\pi(x,t) \neq 0, \quad \forall (x,t) \in \mathbb{R}^d \backslash \{0\} \times [0,\infty), \tag{5.57}$$

which is implied by the assumptions of Th. 37, see Rem. 5.3.1 point (*ii*). The outward unit normal to $\Omega_p(t)$ at x with $p = \pi(x,t)$ is the vector $\nu(x,t)$ defined by (5.25) and we decompose the velocity vector v according to its normal and tangential components to $\Omega_p(t)$ as defined by (5.26).

We now recall the definition of the speed of a surface (or more generally of a co-dimension 1 manifold).

Definition 39. *Consider a time-dependent smooth regular domain $\Omega(t)$ and a point $x \in \partial\Omega(t)$. Then, the speed $w_\perp(x,t)$ of the surface $\partial\Omega(t)$ at x is defined as follows: define $\nu(x,t)$ the outward unit normal to $\partial\Omega(t)$ at x. Then, for t' close to t, the line drawn from x in the direction of $\nu(x,t)$ intersects $\partial\Omega(t')$ at a unique point $X(t')$. Then*

$$w_\perp(x,t) = \big(\frac{d}{dt'}X(t')\big)|_{t'=t} \cdot \nu(x,t). \tag{5.58}$$

In the case of $\Omega_p(t)$, the speed of the surface is given in the following Lemma:

Lemma 40. *Let $\Omega(t) = \Omega_p(t)$. Then the speed of the surface $\Sigma_p(t)$ as defined in Definition 39 is given by*

$$w_\perp(x,t) = -\frac{\partial_t \pi}{|\nabla\pi|}. \tag{5.59}$$

Proof. We can write $\pi(X(t'),t') = p$, for all t' in a small neighbourhood of t, with

$X(t) = x$. Therefore, using (5.25) and (5.58), and since p is independent on time, we get:

$$\begin{aligned} 0 &= \Big(\frac{d}{dt'}(\pi(X(t'),t'))\Big)\Big|_{t'=t} \\ &= \partial_t\pi(x,t) + \big(\frac{d}{dt'}X(t')\big)|_{t'=t}\cdot\nabla\pi(x,t) \\ &= \partial_t\pi(x,t) + \Big(\big(\frac{d}{dt'}X(t')\big)|_{t'=t}\cdot\nu(x,t)\Big)\,|\nabla\pi(x,t)| \\ &= \partial_t\pi(x,t) + w_\perp(x,t)\,|\nabla\pi(x,t)|, \end{aligned} \tag{5.60}$$

which leads to (5.59) and ends the proof of the Lemma. ∎

To define the material velocity, we will need to introduce its flow:

Definition 41. *Given a vector field $v = v(x,t)$ which we assume continuous, bounded and $\mathcal{C}^1$ with respect to x, the flow of v is the unique map $\Phi_t^s : \Omega_N(t) \to \Omega_N(s)$ such that for any $x \in \Omega_N(t)$, the function $\eta : s \mapsto \Phi_t^s(x)$ satisfies*

$$\begin{cases} \eta(t) = x, \\ \eta'(s) = v(\eta(s),s)\ \forall s \geq 0. \end{cases}$$

We can now define the non-swapping constraint for a velocity.

Definition 42. *We assume that the assumptions of Theorem 37 are satisfied. We also assume the non-degeneracy condition (5.57). The material velocity $v(x,t)$ satisfying the same assumptions as in Def. 41 is said to be consistent with the non-swapping constraint if and only if for all (x,t) such that x is a regular point of $W(t,\cdot)$, there exists a neighborhood $\mathcal{U}\times\mathcal{V}\times\mathcal{I}$ of $(x,\pi(t,x),t)$ in $\mathbb{R}^d\times[0,N]\times[0,\infty)$ and a function $H_t : (p,s)\in\mathcal{V}\times\mathcal{I} \mapsto H_t^s(p)\in\mathbb{R}$ which is continuous and $\mathcal{C}^1$ with respect to s, such that for any $s\in\mathcal{I}$ the map $p\in\mathcal{V}\mapsto H_t^s(p)\in\mathbb{R}$ is injective and such that for all $(y,s)\in\mathcal{U}\times\mathcal{I}$, we have*

$$\pi(\Phi_t^s(y),s) = H_t^s(\pi(y,t)). \tag{5.61}$$

Remark 5.4.1. *Def. 42 is illustrated by Fig. 5.5: Eq. (5.61) implies that, if at time t two cells are at neighboring locations y_1 and y_2 (namely y_1 and y_2 belong to the neighborhood $\mathcal{U}$) such that they belong to the same level set, i.e. $p = \pi(y_1,t) = \pi(y_2,t)$ (respectively do not belong to the same level set i.e. $\pi(y_1,t)\neq\pi(y_2,t)$), then at time s they belong to the same level set given by $\pi(\Phi_t^s(y_1),s) = \pi(\Phi_t^s(y_2),s) = H_t^s(p)$ (respectively they do not belong to the same level set i.e. $p = \pi(\Phi_t^s(y_1),s) = H_t^s(\pi(y_1,t)) \neq p' = \pi(\Phi_t^s(y_2),s) = H_t^s(\pi(y_2,t))$ because of the injectivity of H_t^s).*

Remark 5.4.2. *In dimension 1 the non-swapping constraint is always satisfied and therefore carries no content. Indeed, since we suppose in Def. 42 that x is a regular point of*

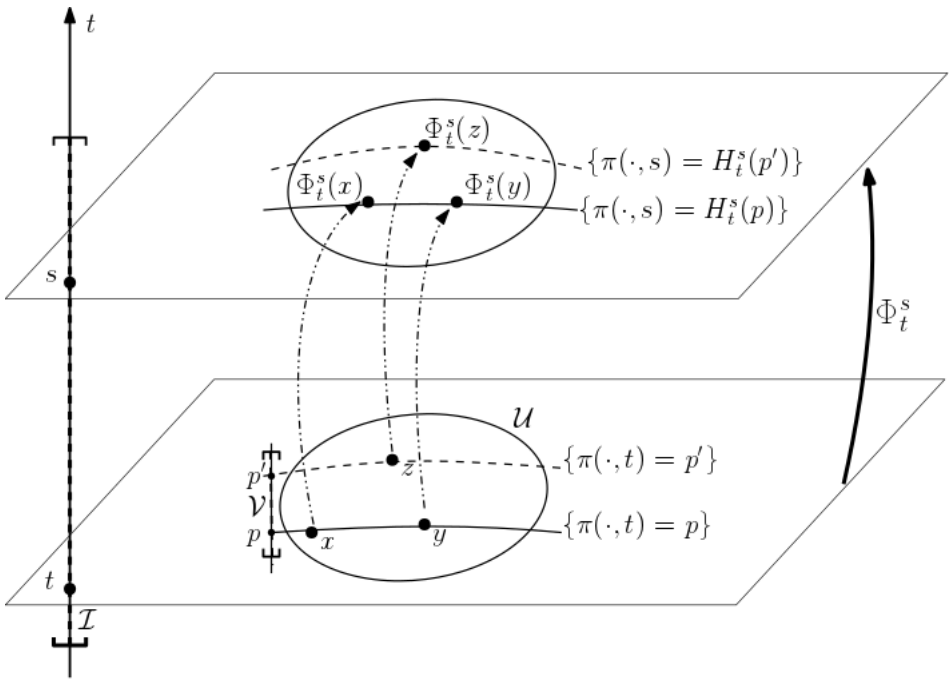


Figure 5.5: Schematics of the non-swapping condition in Def. 42

$W(\cdot,t)$, then it is also a regular point of $\pi(\cdot,t)$, and we can locally invert $\pi(\cdot,t)_{|\mathcal{U}} : \mathcal{U} \to \mathcal{V}$. Thus we can always find a function H_t satisfying Eq. (5.61) as

$$H_t^s(p) := \pi\left(\Phi_t^s \circ \pi_{|\mathcal{U}}(\cdot,t)^{-1}(p), s\right).$$

Next, we give a necessary condition that the velocity v has to fulfil when the evolution of n is given by the continuity equation. Particularly, we show that in dimension $d \geq 2$, if a particle moves with velocity v satisfying the non-swapping constraint, then the normal component of the velocity is given by the domain velocity of its level set $\Sigma_p(t)$ (Prop. 43 below), and it remains in the same level set $\Sigma_p(t)$ for all times (Prop. 45 below). This shows that Definition 42 ensures that a particle remains in the layer between two level sets $\Sigma_p(t)$ and $\Sigma_{p+\delta p}(t)$ at all times. More precisely:

Proposition 43. *Suppose that v satisfies the assumptions of Def. 41, the non-swapping constraint as given by Def. 42 for $d \geq 2$, and is such that*

$$\partial_t n + \nabla \cdot (vn) = 0. \tag{5.62}$$

Then, we have

$$v \cdot \nu = w_\perp, \tag{5.63}$$

where $w_\perp$ is given by Eq. (5.59) and ν by (5.25).

To prove this result, we first show the following lemma that provides a global version of the non-swapping constraint:

Lemma 44. *Let v satisfy the assumptions of Def. 41 and verify the non-swapping constraint as expressed by Def. 42. We assume $d \geq 2$. Then, there exists a continuous function $h = h(p,t)$ such that for all t, x,*

$$(\partial_t + v \cdot \nabla_x)\pi(x,t) = h(\pi(x,t),t). \tag{5.64}$$

Proof. For all (t,x) such that x is not a critical point of $W(\cdot,t)$, the non-swapping constraint in Def. 42 gives a function H_t that satisfies (5.61) for all y in a neighbourhood $\mathcal{U}_x$ of x. Differentiating (5.61) along s and evaluating at $s = t$ we have:

$$(\partial_t + v \cdot \nabla)\pi(y,t) = \partial_s H_t^s(\pi(y,t))|_{s=t}\,,$$

for all $y \in \mathcal{U}_x$. We define $h_x(p,t) = \partial_s H_t^s(\pi(y,t))|_{s=t}$, with $\pi(y,t) = p$. We will show that this definition is independent of x. Indeed, if x, y are in $\Omega_N(t)$ (and are not critical points), and $z \in \mathcal{U}_x \cap \mathcal{U}_y$, then it must hold

$$h_x(\pi(z,t),t) = (\partial_t + v \cdot \nabla)\pi(z,t) = h_y(\pi(z,t),t). \tag{5.65}$$

Now, since $d \geq 2$ and $W(\cdot,t)$ has a unique critical point (at $x = 0$), the level sets of π are diffeomorphic to connected $(d-1)-$spheres. Using the relation (5.65) and the connectivity of the level sets (since $d \geq 2$), we get that $h_x(p,t) = h_y(p,t)$ for any x, y such that $h_x(\cdot,t)$ and $h_y(\cdot,t)$ are defined at p. Thus the functions h_x can be glued to a single function $h = h(p,t)$ that satisfies (5.64). Since the functions h_x are continuous, h is continuous as well. ∎

Proof of Prop. 43. By lemma 44 there exists a function h satisfying Eq. (5.64). This equation is equivalent to

$$\begin{aligned} v \cdot \nu(x,t) &= \frac{h(\pi(x,t),t) - \partial_t \pi(x,t)}{|\nabla \pi(x,t)|} \\ &= \frac{h(\pi(x,t),t)}{|\nabla \pi(x,t)|} + w_\perp(x,t), \end{aligned}$$

where $w_\perp$ is the normal velocity of $\Sigma_p(t)$ as computed in (5.59). Since $\int_{\Omega_p(t)} n \, dx = p$ by

the definition of $\Omega_p(t)$, we deduce that

$$\begin{aligned}
0 &= \frac{d}{dt}\left(\int_{\Omega_p(t)} n\ dx\right) \\
&= \int_{\Omega_p(t)} \partial_t n\ dx + \int_{\Sigma_p(t)} n w_\perp dS(x) \\
&= \int_{\Omega_p(t)} -\nabla\cdot(nv)\ dx + \int_{\Sigma_p(t)} n w_\perp dS(x) \\
&= \int_{\Sigma_p(t)} n(w_\perp - v\cdot\nu) dS(x) \\
&= -h(p,t)\int_{\Sigma_p(t)} \frac{n}{|\nabla\pi|} dS(x).
\end{aligned}$$

In the second line, we used the standard formula for the derivative of an integral on a time-dependent domain. The continuity equation was used in the third line, and Stokes' theorem in the fourth line. Since $n > 0$ on $\Sigma_p(t)$, and since $\Sigma_p(t)$ has positive $d-1$ measure (because p is not a critical value of the potential), the integral on the last line is strictly positive. We conclude that $h(p,t) = 0$ for all $p > 0$, and so for all (t,x), we have

$$0 = (\partial_t + v\cdot\nabla)\pi = \partial_t\pi + (v\cdot\nu)|\nabla\pi|,$$

which is exactly (5.63) and finishes the proof. ∎

As a consequence of the previous proof, we have

Proposition 45. *Suppose n satisfies the continuity equation (5.62), v satisfies the non-swapping constraint as expressed in Def. 42 and $d \geq 2$. Then there exists a constant $p \geq 0$, such that $\Phi_0^t(x) \in \Sigma_p(t)$, $\forall t \geq 0$.*

Proof. Let $p = \pi(x,0)$, it follows from 43 that

$$\frac{d}{dt}\{\pi(\Phi_0^t(x),t)\} = (\partial_t\pi + v\cdot\nabla\pi)(\Phi_0^t(x),t) = 0.$$

And so $\pi(\Phi_0^t(x),t) = p$ for all $t \geq 0$, which proves the proposition. ∎

We now show that a velocity field satisfying the non-swapping condition in dimension $d \geq 2$ (42) satisfies the continuity equation averaged over all surfaces $\Sigma_p(t)$. In other words, the number of particles leaving $\Sigma_p(t)$ at a given time is exactly compensated by the number of particles arriving at $\Sigma_p(t)$.

Theorem 46. *Under the assumptions of Theorem 37, let $n(x,t)$ be given by (5.13). Let v be a vector field such that*

$$v\cdot\nu = w_\perp, \tag{5.66}$$

where ν and $w_\perp$ are given by Eqs. (5.25) and (5.59) respectively. Then, such vector field satisfies

$$\left\langle \delta \circ (\pi(\cdot,t) - p)\,,\, \big(\partial_t n + \nabla\cdot(nv)\big)(\cdot,t) \right\rangle = 0, \quad \forall t > 0, \quad \forall p \in (0,N). \tag{5.67}$$

Remark 5.4.3. *Notice that we exclude the case $p = N$ since then n becomes discontinuous and the derivatives cannot be defined.*

Proof. Note that we have dropped the subscript N to n_N for simplicity. Let $t \geq 0$, since the function

$$p \mapsto \left\langle \delta \circ (\pi(\cdot,t) - p)\,,\, \big(\partial_t n + \nabla\cdot(nv)\big)(\cdot,t) \right\rangle$$

is continuous, we only need to show that for all $p \geq 0$,

$$I(p) := \int_0^p \left\langle \delta \circ (\pi(\cdot,t) - p')\,,\, \big(\partial_t n + \nabla\cdot(nv)\big)(\cdot,t) \right\rangle dp' = 0.$$

Using Stokes' theorem, we have:

$$\begin{aligned} I(p) &= \int_{\{x\,|\,\pi(t,x)\leq p\}} \big(\partial_t n + \nabla\cdot(nv_\perp)\big)(x,t)\,dx \\ &= \int_{\Omega_p(t)} \partial_t n(x,t)\,dx + \int_{\Sigma_p(t)} nv_\perp\cdot\nu(x,t)\,dS(x), \end{aligned}$$

where $dS(x)$ is the canonical measure on $\partial\Omega(t)$. By hypothesis, $v_\perp\cdot\nu(x,t)$ is exactly the velocity of $\Sigma_p(t)$ at (x,t), and so:

$$\begin{aligned} \int_{\Omega_p(t)} \partial_t n(x,t)\,dx + \int_{\Sigma_p(t)} nv_\perp\cdot\nu(x,t)\,dS(x) &= \frac{d}{dt}\left(\int_{\Omega_p(t)} n(x,t)\,dx\right) \\ &= \frac{dp}{dt} \\ &= 0, \end{aligned}$$

where we used Eq. (5.60). So $I(p) = 0$ for all p, which ends the proof. ∎

5.5 Determination of the tangential velocity

In this section, we provide the detailed mathematical discussion of the results summarized in Section 5.2.5.

5.5.1 Dimension one

In this section, we investigate the one-dimensional case. The non-swapping constraint is an empty constraint in this case (see Remark 5.4.2) and there is no tangential velocity. The consequence is that the dynamics of the medium is not governed by the potential (save for the determination of an integration constant), which is an important difference with the higher dimensional case. In dimension one, the continuity equation for n provides a scalar differential equation for the velocity v, which defines it up to a constant, and this constant is determined by the boundary conditions, which indirectly involve the potential, as the following proposition shows.

Proposition 47. *We suppose $d = 1$. Under the assumptions of Theorem 37, there exists a unique velocity v that satisfies the continuity equation (5.62) and which is compatible with n being a solution of the energy minimization problem, given by the conditions*

$$W(a(t), t) = W(b(t), t), \qquad \int_{a(t)}^{b(t)} n(x, t)dx = N, \tag{5.68}$$

where $\Omega(t) = [a(t), b(t)]$. This velocity is given by

$$v(x, t) = \frac{1}{n(x, t)} \left(n(a(t), t)a'(t) - \int_{a(t)}^{x} \partial_t n(y, t)dy \right), \tag{5.69}$$

where $a'(t)$ denotes the time derivative of $a(t)$ and is given by

$$a'(t) = \frac{n(b, t) \left(\partial_t W(b, t) - \partial_t W(a, t) \right) - \partial_x W(b, t) \int_a^b \partial_t n(x, t)\, dx}{n(b, t)\, \partial_x W(a, t) - n(a, t)\, \partial_x W(b, t)}. \tag{5.70}$$

For clarity, the dependence of a and b on t has been dropped. The expression of $b'(t)$, the time derivative of $b(t)$, is given by (5.70) after exchanging a and b.

Proof. The expression of the velocity v is obtained by integrating the continuity equation (5.62) with respect to space on $[a(t), x]$, noting that the velocity at $a(t)$ is precisely $a'(t)$. We just need to verify that the same property is satisfied at $b(t)$, namely that $v(b(t), t) = b'(t)$. Differentiating the second Eq. (5.68) with respect to t gives

$$b'(t)n(b(t), t) - a'(t)n(a(t), t) + \int_{a(t)}^{b(t)} \partial_t n(y, t)dy = 0. \tag{5.71}$$

Using (5.69), this leads to:

$$v(b(t), t) = \frac{1}{n(b(t), t)} \left(n(a(t), t)a'(t) - \int_{a(t)}^{b(t)} \partial_t n(y, t)dy \right) = b'(t),$$

which ends the proof. To find (5.70) we differentiate the first Eq. (5.68) with respect to

t. We find

$$\partial_x W(b(t),t)\, b'(t) - \partial_x W(a(t),t)\, a'(t) + \partial_t W(b(t),t) - \partial_t W(a(t),t) = 0.$$

Together with (5.71), this forms a 2×2 linear system for (a', b') whose solution leads to (5.70) for a' and to the corresponding expression with a and b exchanged for b'. Note that the denominator cannot be 0 as $\partial_x W(a,t)$ and $\partial_x W(b,t)$ have opposite signs and cannot be zero as W has a unique critical point which belongs to the open interval $(a(t), b(t))$. ∎

5.5.2 Dimension $d \geq 2$: tangential velocity is not zero in general

In this section, we show that in dimension $d \geq 2$ in general the velocity field must have a non-zero tangential component $v_\parallel$ to be consistent with the continuity equation. For this purpose, we provide a counter-example in dimension $d = 2$ where the velocity field is defined by $v = w_\perp \nu$ with $w_\perp$ given by (5.63) and which does not fulfil the continuity equation (5.21).

Indeed, consider a potential $V(x,\tau)$ which does not depend on τ and is of the form

$$V(x) = W(x) = \frac{x_2^2}{2} := \tilde{W}(x_2), \qquad \text{for } x = (x_1, x_2) \in \mathbb{T} \times \mathbb{R},$$

and an average volume

$$\tau(x,t) = |x|t, \quad x \in \mathbb{T} \times \mathbb{R}, \quad t \in [0,\infty). \tag{5.72}$$

Here $\mathbb{T} = (-1,1] \approx \mathbb{R}/2\mathbb{Z}$ is the torus, i.e. we assume that all functions are 2-periodic with respect to x_1 and when integrals with respect to x_1 are involved, they are meant over the torus $\mathbb{T}$. Then, by Prop. 37 it holds that

$$n(x,t) = \frac{1}{\tau(x,t)} = \frac{1}{|x|t}, \quad x \in \mathbb{T} \times \mathbb{R}, \quad t \in [0,\infty).$$

Firstly notice that

$$\pi(x,t) = \int_{\{\tilde{W}(y_2) \leq \tilde{W}(x_2)\}} \tau^{-1}(y,t)\, dy := \tilde{\pi}(x_2,t),$$

so it is x_1-independent. The choice of x_1 lying in the torus $\mathbb{T}$ ensures that this integral is finite. Denoting by (e_1, e_2) a cartesian basis associated to the coordinate system (x_1, x_2), we get that $\nu(x,t)$ is parallel to e_2, i.e.

$$\nu(x,t) = e_2 \text{ for } x_2 > 0, \quad \nu(x,t) = -e_2 \text{ for } x_2 < 0.$$

We also have

$$v_\perp(x,t) = -(\partial_t\pi/|\nabla\pi|)(x,t) = -(\partial_t\tilde{\pi}/|\partial_{x_2}\tilde{\pi}|)(x_2,t) := \tilde{v}_\perp(x_2,t),$$

also only depends on x_2.

This implies

$$0 = \partial_t n + \nabla\cdot(nv) = \partial_t n + \partial_{x_2}(n\tilde{v}_\perp).$$

For the considered value of τ in (5.72) and $x_2 > 0$, we have

$$\partial_t n + \partial_{x_2}(n\tilde{v}_\perp) = \frac{|x|^2(-1 + t\,\partial_{x_2}\tilde{v}_\perp(x_2,t)) - \tilde{v}_\perp(x_2)x_2 t}{|x|^3 t^2}.$$

If this last expression was zero, it would imply that

$$\frac{t}{|x|^2} = \frac{-1 + t\,\partial_{x_2}\tilde{v}_\perp(x_2,t)}{\tilde{v}_\perp(x_2,t)\,x_2},$$

but this cannot hold since the left-hand side depends on x_1 but the right-hand side does not. Hence, we must conclude that the continuity equation is not satisfied.

Remark 5.5.1. *The example proposed here does not satisfy the assumptions of Th. 37, however it can be seen as a limiting case of $\tau^\varepsilon(x,t) = (|x|^2+\varepsilon)^{1/2}\,t$ and $V(x) = ((\varepsilon x_1^2) + x_2^2)/2$ as $\varepsilon \to 0$; and where we have replaced assumption (vii) by periodicity conditions in the first component x_1.*

5.5.3 Dimension $d \geq 2$: determination of $v_\parallel$ under the principle of minimal displacements

We first show that (5.31) is a necessary solvability condition for (5.30). This is a consequence of the following lemma, in which we forget the time variable t:

Lemma 48. *Let f: $\mathbb{R}^d \mapsto \mathbb{R}$ be a smooth function, with d a positive integer. If there exists a smooth vector field A: $\mathbb{R}^d \mapsto \mathbb{R}^d$, tangent to all surfaces Σ_p, i.e. satisfying*

$$A\cdot\nabla\pi = 0, \qquad \text{in} \quad \Omega_N, \tag{5.73}$$

and solving the equation

$$\nabla\cdot A = f, \qquad \text{in} \quad \Omega_N, \tag{5.74}$$

then f must be of zero-average on all level sets Σ_p, i.e. f must satisfy (5.31).

Proof. We show that if A: $\mathbb{R}^d \mapsto \mathbb{R}^d$ is a smooth vector field tangent to all surfaces Σ_p, then, it satisfies

$$\left\langle\, \delta\circ\big(\pi - p\big)\,,\, \nabla\cdot A\,\right\rangle = 0, \qquad \forall p\in(0,N]. \tag{5.75}$$

This will show the result as applying (5.75) to (5.74) leads to (5.31). To show (5.75), we take any smooth function g: $\mathbb{R} \mapsto \mathbb{R}$ with compact support and compute, using (5.43) and Green's formula:

$$\begin{aligned}\int_{-\infty}^{\infty} g(p) \left\langle \delta \circ \big(\pi - p\big) \,,\, \nabla \cdot A \right\rangle dp &= \int_{\mathbb{R}^d} (g \circ \pi)(x)\, (\nabla \cdot A)(x)\, dx \\ &= -\int_{\mathbb{R}^d} \nabla (g \circ \pi)(x) \cdot A(x)\, dx \\ &= -\int_{\mathbb{R}^d} (g' \circ \pi)(x)\, (\nabla \pi \cdot A)(x)\, dx \\ &= 0,\end{aligned}$$

where the cancellation comes from (5.73). This shows (5.75) and ends the proof of the lemma. ∎

Next, we consider the resolution of (5.35) and postpone the proof that the solution of problem (5.33) is given by (5.34) to the end of the section. For any $(p,t) \in (0,N) \times (0,\infty)$, we note that $\Sigma_p(t) \subset \Omega_N(t)$. We denote by $\mathcal{I}_{p,t}$: $\Sigma_p(t) \to \Omega_N(t)$ the set injection of $\Sigma_p(t)$ into $\Omega_N(t)$, i.e. for any $y \in \Sigma_p(t)$, $\mathcal{I}_{p,t}(y) = y \in \Omega_N(t)$. Now, we introduce the following change of variables. For a function θ: $(x,t) \in \cup_{t\in(0,\infty)} \Omega_N(t) \times \{t\} \mapsto \theta(x,t) \in \mathbb{R}$, we define a function $\bar{\theta}$: $(p,t,y) \in \cup_{(p,t)\in(0,N)\times(0,\infty)} \{(p,t)\} \times \Sigma_p(t) \mapsto \bar{\theta}(p,t,y) \in (0,\infty)$ such that

$$\theta(\mathcal{I}_{p,t}(y), t) = \bar{\theta}(p,t,y). \tag{5.76}$$

Below, we will use that

$$(\nabla_{\|}\theta)(\mathcal{I}_{p,t}(y), t) = \nabla_y \bar{\theta}(p,t,y), \tag{5.77}$$

where ∇_y denote the gradient operator on the manifold $\Sigma_p(t)$. We now state the

Theorem 49. *Under the assumptions of Theorem 37 and under the solvability condition (5.31), Eq. (5.35) together with the zero-average constraint (5.36) has a unique solution which can be written $\theta(x,t) = \bar{\theta}(p,t,y)$ thanks to the change of variables (5.76), such that $\bar{\theta}$ belongs to the class $C^0\big((0,N)\times(0,\infty), H^1(\Sigma_p(t))\big)$ where $H^1(\Sigma_p(t))$ is the Sobolev space of square integrable functions on $\Sigma_p(t)$ whose first order distributional derivatives are square integrable.*

Proof. Notice that f (given by (5.30)) is smooth, since τ^{-1} and π are smooth (see Assumption (i) in Th. 37 and Rem. 5.3.1 point (ii)). Taking ψ: $(x,t) \in \mathbb{R}^d \mapsto \psi(x,t) \in \mathbb{R}$ any smooth compactly supported function, multiplying (5.35) by ψ and using Green's formula, we get:

$$\int_0^{\infty} \int_{\mathbb{R}^d} \tau^{-1}(x,t)\, \nabla_{\|}\theta(x,t) \cdot \nabla_{\|}\psi(x,t)\, dx\, dt = \int_0^{\infty} \int_{\mathbb{R}^d} f(x,t)\, \psi(x,t)\, dx\, dt,$$

and using (5.41), we deduce:

$$\int_0^\infty \int_0^N \int_{x\in\Sigma_p(t)} \tau^{-1}(x,t)\,\nabla_\parallel\theta(x,t)\cdot\nabla_\parallel\psi(x,t)\,\frac{dS_{p,t}(x)}{|\nabla\pi(x,t)|}\,dp\,dt$$
$$= \int_0^\infty \int_0^N \int_{x\in\Sigma_p(t)} f(x,t)\,\psi(x,t)\,\frac{dS_{p,t}(x)}{|\nabla\pi(x,t)|}\,dp\,dt, \qquad (5.78)$$

where $dS_{p,t}(x)$ is the Euclidean surface measure on $\Sigma_p(t)$. Using the change of variable (5.76) on both θ and ψ, we get

$$\int_0^\infty \int_0^N \int_{y\in\Sigma_p(t)} \tau^{-1}(\mathcal{I}_{p,t}(y),t)\,\nabla_y\bar\theta(p,t,y)\cdot\nabla_y\bar\psi(p,t,y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}\,dp\,dt$$
$$= \int_0^\infty \int_0^N \int_{y\in\Sigma_p(t)} f(\mathcal{I}_{p,t}(y),t)\,\bar\psi(p,t,y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}\,dp\,dt. \qquad (5.79)$$

Since this is true for any function $\bar\psi(p,t,y)$, this implies that for any $(p,t)\in(0,N)\times(0,\infty)$, and any smooth function $\xi : y\in\Sigma_p(t)\mapsto\xi(y)\in\mathbb{R}$, we have

$$\int_{y\in\Sigma_p(t)} \tau^{-1}(\mathcal{I}_{p,t}(y),t)\,\nabla_y\bar\theta(p,t,y)\cdot\nabla_y\xi(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}$$
$$= \int_{y\in\Sigma_p(t)} f(\mathcal{I}_{p,t}(y),t)\,\xi(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}\,. \qquad (5.80)$$

Eq. (5.79) is the weak formulation of an elliptic problem posed on the closed (i.e. without boundary) smooth manifold $\Sigma_p(t)$. Reciprocally, if $y\mapsto\bar\theta(p,t,y)$ is a solution to (5.80) for any $(p,t)\in(0,N)\times(0,\infty)$, then $\theta(x,t)$ constructed through (5.76) is a solution to (5.78) and ultimately to (5.35).

We now show that (5.80) is equivalent to the same problem when we restrict ξ to satisfy the additional constraint

$$\left\langle\,\delta\circ\big(\pi(\cdot,t)-p\big)\,,\,\xi\,\right\rangle = 0,$$

i.e.

$$\int_{y\in\Sigma_p(t)} \xi(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|} = 0. \qquad (5.81)$$

Indeed, if (5.80) is satisfied for all smooth ξ, it is satisfied in particular for those which satisfy the additional constraint (5.81). Conversely, suppose that (5.80) is satisfied for all smooth ξ that satisfy (5.81) and take now a smooth ξ that does not satisfy (5.81). We define

$$\tilde\xi(y) = \xi(y) - \frac{\int_{z\in\Sigma_p(t)}\xi(z)\,\frac{dS_{p,t}(z)}{|\nabla\pi(\mathcal{I}_{p,t}(z),t)|}}{\int_{z\in\Sigma_p(t)}\frac{dS_{p,t}(z)}{|\nabla\pi(\mathcal{I}_{p,t}(z),t)|}}.$$

Then, by (5.80) applied with $\tilde{\xi}$ (which is legitimate since $\tilde{\xi}$ satisfies (5.81)), we get

$$\int_{y\in\Sigma_p(t)} \tau^{-1}(\mathcal{I}_{p,t}(y),t)\,\nabla_y\bar{\theta}(p,t,y)\cdot\nabla_y\tilde{\xi}(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|} = \int_{y\in\Sigma_p(t)} f(\mathcal{I}_{p,t}(y),t)\,\tilde{\xi}(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}\,. \tag{5.82}$$

But since $\tilde{\xi}$ differs from ξ by a constant on $\Sigma_p(t)$, the left-hand side of (5.82) is equal to the same expression with ξ instead of $\tilde{\xi}$. Using the assumption (5.31) that f is of zero-average on $\Sigma_p(t)$, the right-hand side of (5.82) is also equal to the same expression with ξ instead of $\tilde{\xi}$. So, we deduce that (5.80) is satisfied for all smooth ξ, not only those which satisfy (5.81).

So, now, we are left with solving (5.79) for all smooth ξ that satisfy (5.81). It is time to set up functional spaces. We consider the space $L^2(\Sigma_p(t))$ of square integrable functions on $\Sigma_p(t)$ endowed with the norm

$$\|u\|_{L^2(\Sigma_p(t))} = \Big(\int_{y\in\Sigma_p(t)} |u(y)|^2\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}\Big)^{1/2},$$

and the Sobolev space $H^1(\Sigma_p(t))$ of functions u of $L^2(\Sigma_p(t))$ which have first order distributional derivatives $\nabla_y u$ in $L^2(\Sigma_p(t))$, endowed with the norm

$$\|u\|_{H^1(\Sigma_p(t))} = \Big(\|u\|^2_{L^2(\Sigma_p(t))} + \|\nabla_y u\|^2_{L^2(\Sigma_p(t))}\Big)^{1/2}.$$

Finally, we introduce the space $H^1_0(\Sigma_p(t))$ of functions $u\in H^1(\Sigma_p(t))$ which have zero average on $\Sigma_p(t)$ i.e. such that

$$\int_{y\in\Sigma_p(t)} u(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|} = 0.$$

The space $H^1_0(\Sigma_p(t))$ is a closed subspace of $H^1(\Sigma_p(t))$ (because $\Sigma_p(t)$ is compact) and so, is a valid Hilbert space to apply Lax-Milgram theorem. Indeed, defining

$$\begin{aligned} a(\theta,\xi) &= \int_{y\in\Sigma_p(t)} \tau^{-1}(\mathcal{I}_{p,t}(y),t)\,\nabla_y\theta(y)\cdot\nabla_y\xi(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|},\\ \langle L,\xi\rangle &= \int_{y\in\Sigma_p(t)} f(\mathcal{I}_{p,t}(y),t)\,\xi(y)\,\frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}, \end{aligned}$$

the problem of finding a solution of (5.79) for all ξ satisfying (5.81) can be recast in the functional setting:

$$\begin{aligned} &\text{Find } \theta\in H^1_0(\Sigma_p(t)) \text{ such that}\\ &\qquad a(\theta,\xi) = \langle L,\xi\rangle, \quad \forall\xi\in H^1_0(\Sigma_p(t)). \end{aligned} \tag{5.83}$$

It is clear that a and L are respectively a continuous bilinear form and a continuous linear form on $H_0^1(\Sigma_p(t))$. The only missing hypothesis to apply Lax-Milgram theorem is the coercivity of a on $H_0^1(\Sigma_p(t))$. For this, we remark that since τ^{-1} is smooth and positive, and since $\Sigma_p(t)$ is compact, there exists $C > 0$ such that $\tau^{-1}(\mathcal{I}_{p,t}(y), t) \geq C > 0$ for all $y \in \Sigma_p(t)$. Then, for all $\xi \in H_0^1(\Sigma_p(t))$

$$a(\xi, \xi) \geq C \int_{y \in \Sigma_p(t)} |\nabla_y \xi(x)|^2 \frac{dS_{p,t}(y)}{|\nabla \pi(\mathcal{I}_{p,t}(y), t)|} := C\, \tilde{a}(\xi, \xi). \tag{5.84}$$

The quadratic form $\tilde{a}(\xi, \xi)$ at the right-hand side of (5.84) is nothing but the quadratic form associated to the Laplace Beltrami operator on $\Sigma_p(t)$ endowed with the metric $g(y) = |\nabla \pi(\mathcal{I}_{p,t}(y), t)|^{-\frac{2}{d-1}}\, g_e(y)$, where $g_e(y)$ is the Euclidean metric of $\Sigma_p(t)$ at point y. We know from the properties of the Laplace Beltrami operator on closed (i.e. without boundary) manifolds (see [72], Section 4.D.2) that its leading eigenvalue is zero, is simple and that the associated eigenfunctions are the constants. Furthermore, the eigenfunctions of this Laplace-Beltrami operator form a complete ortho-normal basis of the space $L^2(\Sigma_p(t))$. Therefore, from standard spectral theory, since $H_0^1(\Sigma_p(t))$ is the orthogonal space to the constants for the inner product of $L^2(\Sigma_p(t))$, we have

$$\min_{\xi \in H_0^1(\Sigma_p(t))} \frac{\tilde{a}(\xi, \xi)}{\|\xi\|_{L^2(\Sigma_p(t))}} = \lambda_1 > 0,$$

where λ_1 is the first non-zero eigenvalue of the Laplace-Beltrami operator, which is strictly positive. Therefore, we have

$$a(\xi, \xi) \geq C\, \lambda_1 \|\xi\|^2_{L^2(\Sigma_p(t))}, \quad \forall \xi \in H_0^1(\Sigma_p(t)),$$

with $C\lambda_1 > 0$, which shows the coercivity of a. Applying Lax-Milgram's theorem, we deduce that there exists a unique solution to (5.83). Moreover, by the regularity (in H^1) of the solution with respect to the data, and owing to the fact that all data are smooth, we deduce that the solution $\bar{\theta}$ has the regularity $C^0\big((0, N) \times (0, \infty), H^1(\Sigma_p(t))\big)$, which ends the proof of Theorem 49. ∎

We note that if the problem has spherical symmetry, the solution θ has also spherical symmetry, and the level sets $\Sigma_p(t)$ are spheres. Therefore, θ is constant on $\Sigma_p(t)$ but on the other hand, condition (5.36) implies that its average must be zero. Therefore, the constant value of θ on $\Sigma_p(t)$ is necessarily zero. Thus, when the problem has spherical symmetry, the unique solution of (5.35), (5.36) is zero, the tangential velocity $v_\parallel = 0$ and the velocity v is purely normal $v = w_\perp \nu$.

Now we show that the solution of minimization problem (5.33) is given by 5.34). More precisely, we have the following:

Proposition 50. *Let $v_\parallel$ be a solution of (5.33). Then, there exists a function θ such that*

(5.34) holds.

Proof. Suppose $v_\parallel = v_\parallel(x,t)$ is a solution of (5.33). Let $\delta v_\parallel = \delta v_\parallel(x,t)$ be a variation of $v_\parallel$. Then $\delta v_\parallel$ is a tangent vector field to all level surfaces $\Sigma_p(t)$, for all $(p,t) \in (0,N) \times (0,\infty)$ and satisfies the constraint

$$\nabla \cdot \delta v_\parallel = 0, \quad \forall (x,t) \in \bigcup_{t\in(0,\infty)} \Omega_N(t) \times \{t\}. \tag{5.85}$$

Taking smooth functions φ: $(x,t) \in \cup_{t\in(0,\infty)} \Omega_N(t) \times \{t\} \mapsto \varphi(x,t) \in \mathbb{R}$, and g: $p \in (0,N) \mapsto g(p) \in \mathbb{R}$, we have, successively using Green's formula, the fact that $\delta v_\parallel$ is tangent to $\Sigma_p(t)$, and that $\nabla_\parallel(g \circ \pi) = 0$:

$$\begin{aligned}
0 &= \int_{\Omega_N(t)} \nabla \cdot \delta v_\parallel(x,t)\, \varphi(x,t)\, g(\pi(x,t))\, dx \\
&= -\int_{\Omega_N(t)} \delta v_\parallel(x,t) \cdot \nabla(\varphi\ g \circ \pi)(x,t)\, dx \\
&= -\int_{\Omega_N(t)} \delta v_\parallel(x,t) \cdot \nabla_\parallel(\varphi\ g \circ \pi)(x,t)\, dx \\
&= -\int_{\Omega_N(t)} \delta v_\parallel(x,t) \cdot \nabla_\parallel \varphi(x,t)\ g(\pi(x,t))\, dx \\
&= -\int_0^N \big\langle \delta \circ (\pi(\cdot,t) - p), \delta v_\parallel \cdot \nabla_\parallel \varphi \big\rangle\, g(p)\, dp,
\end{aligned}$$

where the last identity follows from (5.43). Now, since this identity is true for all smooth functions $g(p)$, we deduce that

$$0 = \big\langle \delta \circ (\pi(\cdot,t) - p), \delta v_\parallel \cdot \nabla_\parallel \varphi \big\rangle, \quad \forall (p,t) \in (0,N) \times (0,\infty),$$

or, using (5.29) and the change of variables (5.76):

$$0 = \int_{y\in\Sigma_p(t)} \overline{\delta v}_\parallel(p,t,y) \cdot \nabla_y \bar{\varphi}(p,t,y)\, \frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}, \ \forall (p,t) \in (0,N) \times (0,\infty). \tag{5.86}$$

Now, the Euler-Lagrange equations of the minimization problem (5.33) are written

$$\begin{aligned}
\big\langle \delta \circ (\pi(\cdot,t) - p), v_\parallel \cdot \delta v_\parallel(\cdot,t) \big\rangle = 0, \quad &\forall\, \delta v_\parallel \text{ tangent vector field to } \Sigma_p(t) \\
&\text{and satisfying (5.85) }, \ \forall (p,t) \in (0,N) \times (0,\infty),
\end{aligned} \tag{5.87}$$

or, using (5.29) and (5.76) again:

$$\begin{aligned}
0 = \int_{y\in\Sigma_p(t)} \bar{v}_\parallel(p,t,y) \cdot \overline{\delta v}_\parallel(p,t,y) \frac{dS_{p,t}(y)}{|\nabla\pi(\mathcal{I}_{p,t}(y),t)|}, \quad &\forall\, \delta v_\parallel \text{ tangent vector} \\
&\text{field to } \Sigma_p(t) \text{ and satisfying (5.86) }, \ \forall (p,t) \in (0,N) \times (0,\infty),
\end{aligned} \tag{5.88}$$

Eq. (5.88) shows that on each surface $\Sigma_p(t)$, $\bar{v}_\parallel(p,t,\cdot)$ is a tangent vector field orthogonal (for the $L^2(\Sigma_p(t))$ inner product) to all tangent vector fields $\overline{\delta v}_\parallel(p,t,\cdot)$ themselves orthogonal to all gradient vector fields (by (5.86)). But the space of gradients of functions of $H^1(\Sigma_p(t))$ is the same as the space of gradients of functions of $H^1_0(\Sigma_p(t))$. And this latter space is closed in $L^2(\Sigma_p(t))$. This follows easily again from the coercivity of the quadratic form $\tilde{a}$ as proved in the proof of Theorem 49 (details are left to the reader). Therefore, $\bar{v}_\parallel(p,t,\cdot)$ being orthogonal to the orthogonal space to the gradients (and the space of gradients being closed), is itself a gradient. So, there exists a function $\bar{\theta}(p,t,\cdot)$ (parametrized by $(p,t) \in (0,N) \times (0,\infty)$) such that

$$\bar{v}_\parallel(p,t,y) = \nabla_y \bar{\theta}(p,t,y), \quad \forall y \in \Sigma_p(t), \quad \forall (p,t) \in (0,N) \times (0,\infty).$$

Defining $\theta(x,t)$ through the change of variables (5.76), we get (5.34), which ends the proof. ∎

5.6 Conclusions and future work

In this Chapter, we have proposed a new continuum model of a swelling or drying material. Two aspects have been investigated. The first one is an equilibrium problem describing particles seeking to minimize their mechanical energy subject to non-overlapping constraints. Its solution has been fully characterized. The second one is a non-equilibrium problem where we assume that the particle average volume and potential energy may vary with time and where we compute the resulting velocity applying two principles: (i) the non swapping condition and (ii) the principle of smallest displacements. Under these two principles, the medium velocity has been fully determined. A detailed discussion has been provided and many different elaborations of the model have been proposed. In future work, we intend to progress towards the resolution of the many open problems outlined at the end of Sec. 5.2.6.

6

Conclusion and perspectives

In this Chapter we present the general conclusions of this thesis and we motivate several research directions that can be further pursued.

6.1 Conclusion

Large and congested many-particle systems encompass complex multiscale phenomena, ranging from contacts between neighbouring particles to the emergence of coordinated macroscopic behaviour. These phenomena present big challenges to the modelling and simulation of such systems. In this thesis we tackle some of the challenges and we develop new modelling frameworks and efficient computational tools to study these systems.

Specifically, two modelling frameworks are developed. The first framework describes the system at the particle level, leading to particle-based models, which are studied in Part I and Part II. The second framework describes the system at the number density level, leading to a continuum model, which is described in Part III. The long term aim is to obtain an integrated multiscale modelling framework which is able to capture phenomena occurring across different time and space scales. These phenomena are especially relevant in certain particle systems arising in biology (see [29] for a review). One example is the emergence of the shape of organs during embryonic development from local mechanical interactions occurring between neighbouring cells [118].

The two frameworks are based on the same modelling principles. The system is first assumed to be at a mechanical equilibrium, described by a minimizer of a confining potential subject to volume exclusion constraints. The system then evolves driven by changes in the potential and in the particle properties. These modelling principles have been tested in practice and they have proved appropriate in the case of a densely packed cell tissue, which is studied in Chapter 4. Among other aspects, the two modelling frameworks differ

from each other with respect to the minimization problem involved in the description of the statics equilibrium. Indeed, in the discrete model, this problem is typically non-convex, a solution is not unique and one has to resort to computational tools to find an approximation of a solution. The numerical and analytical treatment of these problems are discussed in Part I of this thesis. In the continuum model, the minimization problem is convex, a solution is unique and we are able to obtain it analytically. This is presented in Part III. Another difference between the two frameworks is related to the type of information each one provides. In the particle-based framework, we get information about the approximated positions and velocities of each particle over time. In the continuum framework, we obtain the exact averaged positions and velocities of the particles over time and the exact averaged behaviour of the whole population.

We conclude that the particle-based framework provides more detailed spatial information, but it requires the development of sophisticated computational tools to be able to obtain this information in the case of large systems. On the other hand, the continuum framework only provides information about the averaged behaviour of the system, but it may be treated analytically by using available tools, which allows a more accurate analysis of the averaged dynamics. This shows the importance of the two approaches and how they complement each other.

6.2 Perspectives

Besides searching for a mathematical link between the particle-based and continuum frameworks, as referred above, other future research directions can be envisioned. For example the improvement of the algorithms in terms of computational speed and accuracy, the analysis of the models and algorithms or the development of more realistic models by considering three spatial dimensions, more accurate particle shapes or more complex interaction rules. Next we describe how the results of each Chapter could be improved following the lines described above and additionally, how they could be extended to solve problems in social sciences, cell biology, ecology, material sciences, astronomy, meteorology and geometry.

Chapter 1 Following the results presented in Chapter 1, one could investigate what types of packings can be generated by the damped Arrow-Hurwicz algorithm (DAHA) due to the non-uniqueness of local minima and how each packing relates with the parameters and initial configuration. The results of such a study could be compared with existing packing generating algorithms (see [51] for a brief review) to assess the efficiency and the ability of the DAHA to generate different types of packings. Regarding the shape of the configurations, we expect the radial symmetry of the initial configuration to be preserved by the algorithm. In two dimensions we expect to obtain nearly hexagonal packings (see Figure 1.7, Chapter 1) for all choices of parameters and initial configurations for which the

algorithm converges. In the more intriguing case of three spatial dimensions the structure of the resulting configuration is not so obvious. The results of our numerical experiments suggest that the DAHA is able to generate random close packings with a volume fraction of around 64%. However other choices of parameters and initial conditions might lead to other types of packings with higher or lower volume fractions.

After this study, the DAHA could be used to investigate packings of interest, such as random loose and random close packings [137, 164, 166], mono- and polydisperse packings, etc. One could efficiently study several properties of large packings such as the pore-size distribution, packing entropy, number of different packing configurations for a given density, crystaline and glassy states, distribution of number of neighbours, degree of heterogeneity and disorder, short and long range correlations between particles and so on [18, 100, 102]. In geometry, provided the DAHA converges in high dimensions $d > 4$, one could use this algorithm to explore the existence of dense configurations or to find bounds for the kissing number, i.e., the maximum number of equally-sized spheres that can be in contact with the same sphere without overlapping. In dimensions 1 to 4 the kissing number has been proven to be 2, 6, 12 [38] and 24 [133], respectively. For higher dimensions the kissing number is mostly unkown. Finally, as shown in Part II the DAHA can be integrated in time-stepping schemes to obtain algorithms that simulate the dynamics of large and dense particle systems.

Chapter 2 The study developed in Chapter 2 could be attempted in the case of more general convex and non-convex potentials, as well as, in the case of higher spatial dimensions ($d \geq 2$). However a new approach would have to be developed as the one used here relies on the dimension ($d = 1$) and on the shape of the potential. It would be particularly interesting to investigate whether these results also apply to potentials that are commonly used in other applications, such as, gravitational potential, Coulomb potential, chemical and electrical potentials or "social potentials" associated with dissatisfaction or preferences of the individuals [61].

Chapter 3 The results presented in Chapter 3 suggest that the fastest algorithm to simulate ballistic aggregation should result from the combination of a time-stepping with an event-driven scheme. The combined scheme would consist of using the time-stepping scheme during a first part of the dynamics and then switching to the event-driven one. An easy to check criterion to switch between the two schemes should be identified. After the construction of such a scheme, one could consider the full physical system presented in Section 3.2.1 and explore the statistics associated to its dynamics, such as distribution of cluster sizes and velocities or the shape and morphology of the growing clusters. Moreover, since the algorithm is able to deal with very large systems (with 1 million particles), it provides an appropriate tool to explore the relation between particle models and the corresponding kinetic models that are obtained when the number of particles goes to

infinity [120], such as the widely studied Smoluschovsky coagulation equation [6]. One could then use this knowledge to study multiscale aggregation phenomena in physics, such as thin films, clouds, cosmic dust or planet formation [78, 120, 131, 136], or in biology, such as bacteria colony growth or animals aggregation [107, 141].

Chapter 4 The mechanical model for a cell tissue developed in Chapter 4 is discrete in time. In order to both simplify the analysis and provide a more realistic description of the system one could try to write a corresponding time-continuous dynamics. In particular, such a description would allow to study the time-asymptotic behaviour of the system, which would provide criteria to classify different parameter regions. One possible way to formulate such a time-continuous model may be obtained by projecting the solution into the admissible space according to an appropriate norm. The theoretical background to formulate and analyse such a time-continuous model may be found for example in [160]. On the other hand, in order to make the model more realistic, one could extend it to three spatial dimensions and consider more accurate particle representations, such as ellipsoids or polyhedrons. One may also incorporate other elements, such as fibers [149], chemical signals or other subcellular components. As discussed at the end of Chapter 4, this model can be used in the study of embryonic development and in particular to understand how organs acquire its shape and how it remains stable [83]. Also, in the context of cancer, this model can be very useful to study how defects in individual cells may lead to a disruption in the tissue [98, 172]. Such a scenario is particularly challenging to reproduce in the lab due to the lack of tools and technology to induce defects in targeted cells. The model could therefore constitute a very valuable tool in such a study.

Chapter 5 In Chapter 5, a continuum model for swelling materials has been developed. The evolution of the number density of the material is driven by the evolution of an external potential and changes in particle size. We have shown that the component of the medium velocity that is tangent to the level sets of the potential is in general nonzero. This allows the particles to move along regions of constant energy. In the future, these results should be discussed in view of commonly used models that do not incorporate this feature, such as Hele-Shaw type models [146]. In a further future the model could be developed to include cell division or other phenomena of interest and it could be applied to the study of practical situations, such as for example tissue growth or the study of different types of cancer therapeutics.

A

Explicit solutions for the primal and dual problems in dimension $d = 2$

In this appendix we present analytical solutions in dimension $d = 2$ to the primal problem (2.33) and to the corresponding dual problem (2.34) presented in Section 2.4.1 for the Lagrangian $\mathcal{L}$ (2.9) presented in Section 2.2.1. In the cases $N = 2$ and $N = 3$ one can find explicit solutions for both problems. Recall that the solution to the primal problem $\overline{X} = (\overline{\mathbf{x}}_i)_{i=1,\dots,N}$, $\overline{\mathbf{x}}_i \in \mathbb{R}^2$, and the associated Lagrange multipliers, $\overline{\lambda} = (\overline{\lambda}_{ij})_{i<j}$, $\overline{\lambda}_{ij} \in \mathbb{R}_+$, represent a critical point to the Lagrangian, therefore they satisfy the complementary slackness condition (2.6) and the Euler-Lagrange equation (2.7). These two conditions together with primal and dual feasibility, i.e., $F(\bar{X}) \leq 0$ and $\bar{\lambda} \geq 0$, respectively, constitute the necessary conditions for optimality, known as KKT-conditions.

For $N = 2$ in dimension $d = 2$, we get from the KKT-conditions:

$$\begin{cases} (1/N - 2\lambda)(\mathbf{x}_1 - \mathbf{x}_2) = 0 \\ (4R^2 = |\mathbf{x}_1 - \mathbf{x}_2|^2 \text{ and } \lambda \geq 0) \text{ or } (4R^2 < |\mathbf{x}_1 - \mathbf{x}_2|^2 \text{ and } \lambda = 0) \end{cases}$$

Any solution to this system satisfies $2R = |\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|$ and $\bar{\lambda} = 1/4$. The solution to the dual problem is also $\hat{\lambda} = 1/4$. In this case, we have $d^* = p^* = R^2$ and the critical-point $(\bar{X}, \bar{\lambda})$ is a saddle-point of the Lagrangian. Furthermore, the matrix $A(\bar{\lambda})$ is null, therefore the Arrow-Hurwicz method (2.13) does **not** converge towards the saddle-point.

For $N = 3$ in dimension $d = 2$, we obtain:

$$\begin{cases} \mathbf{x}_1 - \langle \mathbf{x} \rangle = 2\big(\lambda_{12}(\mathbf{x}_1 - \mathbf{x}_2) + \lambda_{13}(\mathbf{x}_1 - \mathbf{x}_3)\big) \\ \mathbf{x}_2 - \langle \mathbf{x} \rangle = 2\big(\lambda_{12}(\mathbf{x}_2 - \mathbf{x}_1) + \lambda_{23}(\mathbf{x}_2 - \mathbf{x}_3)\big) \\ \mathbf{x}_3 - \langle \mathbf{x} \rangle = 2\big(\lambda_{13}(\mathbf{x}_3 - \mathbf{x}_1) + \lambda_{23}(\mathbf{x}_3 - \mathbf{x}_2)\big) \\ (4R^2 = |\mathbf{x}_1 - \mathbf{x}_2|^2 \text{ and } \lambda_{12} \geq 0) \text{ or } (4R^2 < |\mathbf{x}_1 - \mathbf{x}_2|^2 \text{ and } \lambda_{12} = 0) \\ (4R^2 = |\mathbf{x}_1 - \mathbf{x}_3|^2 \text{ and } \lambda_{13} \geq 0) \text{ or } (4R^2 < |\mathbf{x}_1 - \mathbf{x}_3|^2 \text{ and } \lambda_{13} = 0) \\ (4R^2 = |\mathbf{x}_2 - \mathbf{x}_3|^2 \text{ and } \lambda_{23} \geq 0) \text{ or } (4R^2 < |\mathbf{x}_2 - \mathbf{x}_3|^2 \text{ and } \lambda_{23} = 0) \end{cases}$$

Apart from permutations and translations, this system has several solutions, including a saddle-point of the Lagrangian, as we will see next. By symmetry, we have $\lambda_{12} = \lambda_{13} = \lambda_{23} = \lambda$. Moreover, from basic triangle geometry we have:

$$\mathbf{x}_1 - \langle \mathbf{x} \rangle = 2\lambda\big((\mathbf{x}_1 - \mathbf{x}_2) + (\mathbf{x}_1 - \mathbf{x}_3)\big) = 2\lambda \cdot 3\big(\mathbf{x}_1 - \langle \mathbf{x} \rangle\big).$$

Thus, all pairs $(\bar{X}, \bar{\lambda})$ satisfying $|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2| = |\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_3| = |\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_3| = 2R$, $\bar{\lambda} = \frac{1}{6}$ are solutions to this system. The solution to the dual problem satisfies $\hat{\sigma}_i = 2/6,\ i = 1, 2, 3$, therefore, $\hat{\lambda}_{12} = \lambda_{13} = \lambda_{23} = 1/6$ is a solution to the dual problem. In this case, we have $d^* = p^* = 2R^2$ and the critical-point $(\bar{X}, \bar{\lambda})$ is a saddle-point of the Lagrangian. Furthermore, the matrix $A(\bar{\lambda})$ is null, therefore the Arrow-Hurwicz method does **not** converge.

For $N = 4$, we look for solutions of the form:

$$\begin{aligned} |\mathbf{x}_1 - \mathbf{x}_2| = |\mathbf{x}_2 - \mathbf{x}_3| = |\mathbf{x}_3 - \mathbf{x}_4| = |\mathbf{x}_4 - \mathbf{x}_1| &= 2R \\ \lambda_{12} = \lambda_{23} = \lambda_{34} = \lambda_{14} &=: \lambda, \\ \lambda_{13} = \lambda_{24} &= 0 \end{aligned}$$

We only need to compute λ. By using a geometrical argument we obtain:

$$\mathbf{x}_1 - \langle \mathbf{x} \rangle = 2\lambda\big((\mathbf{x}_1 - \mathbf{x}_2) + (\mathbf{x}_1 - \mathbf{x}_4)\big) = \lambda\big((\mathbf{x}_1 - \mathbf{x}_2) + (\mathbf{x}_1 - \mathbf{x}_4) + (\mathbf{x}_1 - \mathbf{x}_3)\big) = 4\lambda\big(\mathbf{x}_1 - \langle \mathbf{x} \rangle\big),$$

therefore $\lambda = 1/4$. These are solutions to the primal problem and we get $p^* = 4R^2$. On the other hand, the solution to the dual satisfies $\hat{\sigma}_i = 3/8$ and we have $d^* = 3R^2$, which is strictly smaller than p^*, consequently, a solution of the primal is **not** a saddle-point of the Lagrangian.

References

[1] E. Aarts and J. Korst. *Simulated annealing and Boltzmann machines*. John Wiley and Sons Inc, New York, 1988.

[2] J. Abadie. On the kuhn-tucker theorem. Technical report, DTIC Document, 1966.

[3] V. Acary. Projected event-capturing time-stepping schemes for nonsmooth mechanical systems with unilateral contact and coulomb's friction. *Computer Methods in Applied Mechanics and Engineering*, 256:224–250, 2013.

[4] B. Addis, M. Locatelli, and F. Schoen. Disk packing in a square: a new global optimization approach. *INFORMS Journal on Computing*, 20(4):516–524, 2008.

[5] M. Alam and S. Luding. Rheology of bidisperse granular mixtures via event-driven simulations. *Journal of Fluid Mechanics*, 476:69–103, 2003.

[6] A. Aldous. *Deterministic and stochastic models for coalescence (aggregation and coagolation): a review of the mean-field theory for probabilists*. Bernoulli 5(1), p. 3-48, 1999.

[7] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Oxford university press, 2017.

[8] S. Alt, P. Ganguly, and G. Salbreux. Vertex models: from cell mechanics to tissue morphogenesis. *Phil. Trans. R. Soc. B*, 372(1720):20150520, 2017.

[9] F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1-2):3–11, 2001.

[10] F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with hessian-driven damping.: Application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–779, 2002.

[11] M. B. Amar and P. Ciarletta. Swelling instability of surface-attached gels as a model of soft tissue growth under geometric constraints. *J. Mech. Phys. Solids*, 58(7):935–954, 2010.

[12] D. Ambrosi and L. Preziosi. On the closure of mass balance models for tumor growth. *Math. Models Methods Appl. Sci.*, 12(05):737–754, 2002.

[13] C. Amitrano, A. Coniglio, and F. Di Liberto. Growth probability distribution in kinetic aggregation processes. *Physical review letters*, 57(8):1016, 1986.

[14] A. Anderson and K. Rejniak. *Single-cell-based models in biology and medicine*. Springer Science & Business Media, 2007.

[15] I. S. Aranson and L. S. Tsimring. Patterns and collective behavior in granular media: Theoretical concepts. *Rev. Modern Phys.*, 78(2):641, 2006.

[16] V. I. Arnold. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013.

[17] K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and nonlinear programming*. Stanford University Press, 1958.

[18] V. Baranau, D. Hlushkou, S. Khirevich, and U. Tallarek. Pore-size entropy of random hard-sphere packings. *Soft Matter*, 9(12):3361–3372, 2013.

[19] N. Bellomo, A. Bellouquid, J. Nieto, and J. Soler. Multiscale biological tissue models and flux-limited chemotaxis for multicellular growing systems. *Mathematical Models and Methods in Applied Sciences*, 20(07):1179–1207, 2010.

[20] M. Bernardo, C. Budd, A. R. Champneys, and P. Kowalczyk. *Piecewise-Smooth Dynamical Systems: Theory and Applications*, volume 163. Springer Science & Business Media, 2008.

[21] M. Bertsch, R. Dal Passo, and M. Mimura. A free boundary problem arising in a simplified tumour growth model of contact inhibition. *Interfaces Free Bound.*, 12(2):235–250, 2010.

[22] M. Bertsch, D. Hilhorst, H. Izuhara, and M. Mimura. A nonlinear parabolic-hyperbolic system for contact inhibition of cell-growth. *Differ. Equ. Appl*, 4(1):137–157, 2012.

[23] D. Bertsekas. *Constrainted optimization and Lagrange multiplier methods*. Athena Scientific, Belmont, Massachusetts, 1996.

[24] D. P. Bertsekas. *Nonlinear programming*. Athena scientific, second edition, 1999.

[25] E. G. Birgin, J. M. Martinez, and D. P. Ronconi. Optimizing the packing of cylinders into a rectangular container: a non-linear approach. *Eur. J. Oper. Res.*, 160:19–33, 2005.

[26] M. Bock, A. K. Tyagi, J.-U. Kreft, and W. Alt. Generalized voronoi tessellation as a model of two-dimensional cell tissue dynamics. *Bulletin of mathematical biology*, 72(7):1696–1731, 2010.

[27] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[28] B. Brogliato, A. Ten Dam, L. Paoli, F. Genot, and M. Abadie. Numerical simulation of finite dimensional multibody nonsmooth mechanical systems. *Applied Mechanics Reviews*, 55(2):107–150, 2002.

[29] M. J. Buehler. Multiscale mechanics of biological and biologically inspired materials and structures. *Acta Mechanica Solida Sinica*, 23(6):471–483, 2010.

[30] H. Byrne and D. Drasdo. Individual-based and continuum models of growing cell populations: a comparison. *J. Math. Biol.*, 58(4):657–687, 2009.

[31] H. Byrne and L. Preziosi. Modelling solid tumour growth using the theory of mixtures. *Math. Med. Biol.*, 20(4):341–366, 2003.

[32] M. A. Chaplain and B. Sleeman. A mathematical model for the growth and classification of a solid tumor: a new approach via nonlinear elasticity theory using strain-energy functions. *Math. Biosci.*, 111(2):169–215, 1992.

[33] A. Chauvière, L. Preziosi, and C. Verdier. *Cell mechanics: from single scale-based models to multiscale modeling*. CRC Press, 2010.

[34] C. Chicone. *Ordinary differential equations with applications*. Springer-Verlag New York, Inc., 1999.

[35] F. Clarke. *Functional analysis, calculus of variations and optimal control*, volume 264. Springer Science & Business Media, 2013.

[36] H. Cohn, A. Kumar, S. D. Miller, D. Radchenko, and M. Viazovska. The sphere packing problem in dimension 24. *arXiv preprint arXiv:1603.06518*, 2016.

[37] P. Colli, G. Gilardi, E. Rocca, and J. Sprekels. Vanishing viscosities and error estimate for a Cahn–Hilliard type phase field system related to tumor growth. *Nonlinear Anal. Real World Appl.*, 26:93–108, 2015.

[38] J. Conway and N. Sloane. *Shpere packing, lattices and groups*. Springer, 2nd edition, New York, 1992.

[39] O. Córdoba-Rodríguez, M. del Castillo-Mussot, and J. Montemayor-Aldrete. Some studies on safe maximum packing of live agents in crowds or containers. *Revista mexicana de física*, 55(6):450–455, 2009.

[40] V. Cristini, J. Lowengrub, and Q. Nie. Nonlinear simulation of tumor growth. *J. Math. Biol.*, 46(3):191–224, 2003.

[41] P. A. Cundall and O. D. Strack. A discrete numerical model for granular assemblies. *Geotechnique*, 29(1):47–65, 1979.

[42] J. O. Dada and P. Mendes. Multi-scale modelling and simulation in systems biology. *Integrative Biology*, 3(2):86–96, 2011.

[43] D. Debnath, J. S. Gainer, C. Kilic, D. Kim, K. T. Matchev, and Y.-P. Yang. Identifying phase-space boundaries with voronoi tessellations. *The European Physical Journal C*, 76(11):645, 2016.

[44] P. Degond. Macroscopic limits of the boltzmann equation: a review. In *Modeling and Computational Methods for Kinetic Equations*, pages 3–57. Springer, 2004.

[45] P. Degond, M. A. Ferreira, and S. Motsch. Damped Arrow–Hurwicz algorithm for sphere packing. *J. Comput. Phys.*, 332:47–65, 2017.

[46] P. Degond, P. Minakowski, and E. Zatorska. Transport of congestion in two-phase compressible/incompressible flows. *Nonlinear Analysis: Real World Applications*, 42:485 – 510, 2018.

[47] P. Degond, L. Navoret, R. Bon, and D. Sanchez. Congestion in a macroscopic model of self-driven particles modeling gregariousness. *Journal of Statistical Physics*, 138(1-3):85–125, 2010.

[48] S. Dey, D. Das, and R. Rajesh. Lattice models for ballistic aggregation in one dimension. *EPL (Europhysics Letters)*, 93(4):44001, 2011.

[49] F. L. DiMaggio and I. S. Sandler. Material model for granular soils. *Journal of Engineering Mechanics*, 1971.

[50] J. Dirkse and J. Cawley. A modified ballistic aggregation model. *Journal of colloid and interface science*, 170(2):466–476, 1995.

[51] A. Donev. *Jammed Packing of Hard Particles*. PhD thesis, Princeton University, 2006.

[52] A. Donev, S. Torquato, and F. Stillinger. Neighbor list collision-driven molecular dynamics simulation for nonspherical hard particles. I. Algorithmic details. *J. Comput. Phys.*, 202(2):737–764, 2005.

[53] J. P. Doye. Physical perspectives on the global optimization of atomic clusters. In J. D. Pintér, editor, *Global Optimization*, pages 103–139. Springer, USA, 2006.

[54] D. Drasdo and S. Höhme. A single-cell-based model of tumor growth in vitro: monolayers and spheroids. *Phys. Biol.*, 2(3):133, 2005.

[55] J. Duran. *Sands, powders, and grains: an introduction to the physics of granular materials*. Springer Science & Business Media, 2012.

[56] M. Elimelech, J. Gregory, X. Jia, and R. Williams. *Particle Deposition and Aggregation - Measurement, Modelling and Simulation*. Elsevier, 1995.

[57] M. Escobedo, P. Laurençot, S. Mischler, and B. Perthame. Gelation and mass conservation in coagulation-fragmentation models. *J. Diff. Eq.*, 195(1):143–174, 2003.

[58] M. Escobedo and S. Mischler. Scalings for a ballistic aggregation equation. *Journal of Statistical Physics*, 141(3):422–458, 2010.

[59] R. Farhadifar, J.-C. Röper, B. Aigouy, S. Eaton, and F. Jülicher. The influence of cell mechanics, cell-cell interactions, and proliferation on epithelial packing. *Current Biology*, 17(24):2095–2104, 2007.

[60] S. Faure, A. Lefebvre-Lepot, and B. Semin. Dynamic numerical investigation of random packing for spherical and nonconvex particles. In M. Ismail, B. Maury, and G. J.-F., editors, *ESAIM: Proceedings*, volume 28, pages 13–32. EDP Sciences, 2009.

[61] S. Faure and B. Maury. Crowd motion from the granular standpoint. *Mathematical Models and Methods in Applied Sciences*, 25(03):463–493, 2015.

[62] H. Federer. *Geometric measure theory*. Springer, 2014.

[63] S. J. Ferreira, S. Alves, A. Brito, and J. Moreira. *Morphological transition between diffusion-limited and ballistics aggregation growth patterns*. Physical review E, 71, 5, 2005.

[64] R. A. Fisher. The wave of advance of advantageous genes. *Annals of Human Genetics*, 7(4):355–369, 1937.

[65] A. G. Fletcher, M. Osterfield, R. E. Baker, and S. Y. Shvartsman. Vertex models of epithelial morphogenesis. *Biophysical journal*, 106(11):2291–2304, 2014.

[66] Y. Forterre and O. Pouliquen. Flows of dense granular media. *Annu. Rev. Fluid Mech.*, 40:1–24, 2008.

[67] N. Fournier and P. Laurencot. *Markus-Lushnikov processes, Smoluchowski's and Flory's models*. Stochastic Processes and their Applications, 119, 1, p. 167-189, 2009.

[68] W. Fox. A study of bulk density and water in a swelling soil. *Soil Science*, 98(5):307–316, 1964.

[69] L. Frachebourg. Exact solution of the one-dimensional ballistic aggregation. *Physical review letters*, 82(7):1502, 1999.

[70] L. Frachebourg, P. A. Martin, and J. Piasecki. Ballistic aggregation: a solvable model of irreversible many particles dynamics. *Physica A: Statistical Mechanics and its Applications*, 279(1):69–99, 2000.

[71] A. Friedman. A free boundary problem for a coupled system of elliptic, hyperbolic, and Stokes equations modeling tumor growth. *Interfaces Free Bound.*, 8(2):247–261, 2006.

[72] S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian geometry*, volume 3. Springer, 1990.

[73] F. Ganji, S. Vasheghani-Farahani, and E. Vasheghani-Farahani. Theoretical description of hydrogel swelling: a review. *Iranian Polymer Journal*, 19(5):375–398, 2010.

[74] M. Goodman and S. Cowin. A continuum theory for granular materials. *Arch. Ration. Mech. Anal.*, 44(4):249–266, 1972.

[75] N. R. Goold, E. Somfai, and R. C. Ball. Anisotropic diffusion limited aggregation in three dimensions: Universality and nonuniversality. *Phys. Rev. E*, 72:031403, Sep 2005.

[76] A. Gorban and I. Karlin. Hilbert's 6th problem: exact and approximate hydrodynamic manifolds for kinetic equations. *Bulletin of the American Mathematical Society*, 51(2):187–246, 2014.

[77] R. Grima and S. Schnell. A systematic investigation of the rate laws valid in intracellular environments. *Biophysical chemistry*, 124(1):1–10, 2006.

[78] M. Grzegorczyk, K. Maruszewski, and M. Rybaczuk. Modeling of silica sol–gel thin films with ballistic aggregation. *Chaos, Solitons & Fractals*, 24(4):1053–1058, 2005.

[79] M. Grzegorczyk, M. Rybaczuk, and K. Maruszewski. Ballistic aggregation: an alternative approach to modeling of silica sol–gel structures. *Chaos, Solitons & Fractals*, 19(4):1003–1011, 2004.

[80] C. Guillot and T. Lecuit. Mechanics of epithelial tissue homeostasis and morphogenesis. *Science*, 340(6137):1185–1189, 2013.

[81] J. M. Haile. *Molecular Dynamics Simulation: Elementary Methods*. John Wiley & Sons, Inc. New York, NY, USA, 1992.

[82] T. C. Hales. A proof of the kepler conjecture. *Annals of mathematics*, 162(3):1065–1185, 2005.

[83] E. Hannezo, J. Prost, and J.-F. Joanny. Theory of epithelial sheet morphology in three dimensions. *Proceedings of the National Academy of Sciences*, 111(1):27–32, 2014.

[84] E. Hannezo, J. Prost, and J.-F. Joanny. Theory of epithelial sheet morphology in three dimensions. *Proceedings of the National Academy of Sciences*, 111(1):27–32, 2014.

[85] J.-P. Hansen and I. R. McDonald. *Theory of simple liquids*. Academic Press, New York, second edition, 1990.

[86] B. Hartke. Efficient global geometry optimization of atomic and molecular clusters. In J. D. Pintér, editor, *Global optimization*, pages 141–168. Springer, USA, 2006.

[87] A. Hawkins-Daarud, K. G. van der Zee, and J. Tinsley Oden. Numerical simulation of a thermodynamically consistent four-species tumor growth model. *Int. J. Numer. Methods Biomed. Eng.*, 28(1):3–24, 2012.

[88] S. Hecht and N. Vauchelet. Incompressible limit of a mechanical model for tissue growth with non-overlapping constraint. *arXiv preprint arXiv:1702.08850*, 2017.

[89] M. Hifi and R. M'hallah. A literature review on circle and sphere packing problems: models and methodologies. *Adv. Oper. Res.*, 2009:Article ID 150624, 2009.

[90] D. Hilhorst, J. Kampmann, T. N. Nguyen, and K. G. Van Der Zee. Formal asymptotic limit of a diffuse-interface tumor-growth model. *Math. Models Methods Appl. Sci.*, 25(06):1011–1043, 2015.

[91] A. Hočevar Brezavšček, M. Rauzi, M. Leptin, and P. Ziherl. A Model of Epithelial Invagination Driven by Collective Mechanics of Identical Cells. *Biophysical Journal*, 103(5):1069–1077, Sept. 2012.

[92] S. Ishihara, P. Marcq, and K. Sugimura. From cells to tissue: A continuum model of epithelial mechanics. *Physical Review E*, 96(2):022418, 2017.

[93] M. Jarvis, E. Mackenzie, and H. Duncan. The textural analysis of cooked potato. 2. swelling pressure of starch during gelatinisation. *Potato Research*, 35(1):93–102, 1992.

[94] Y. Jiang and F. Leyvraz. Kinetic properties of ballistic aggregation. *Physical Review E*, 50(3):2148, 1994.

[95] W. Jin, C. J. Penington, S. W. McCue, and M. J. Simpson. Stochastic simulation tools and continuum models for describing two-dimensional collective cell spreading with universal growth functions. *Physical biology*, 13(5):056003, 2016.

[96] S. T. Johnston, R. E. Baker, D. S. McElwain, and M. J. Simpson. Co-operation, competition and crowding: a discrete framework linking allee kinetics, nonlinear diffusion, shocks and sharp-fronted travelling waves. *Scientific reports*, 7:42134, 2017.

[97] A. Jones, H. Byrne, J. Gibson, and J. Dold. A mathematical model of the stress induced during avascular tumour growth. *Journal of mathematical biology*, 40(6):473–499, 2000.

[98] K. H. Kain, J. W. Miller, C. R. Jones-Paris, R. T. Thomason, J. D. Lewis, D. M. Bader, J. V. Barnett, and A. Zijlstra. The chick embryo as an expanding experimental model for cancer and cardiovascular research. *Developmental Dynamics*, 243(2):216–228, 2014.

[99] M. Kardar, G. Parisi, and Y.-C. Zhang. Dynamic scaling of growing interfaces. *Physical Review Letters*, 56(9):889, 1986.

[100] S. Khirevich, A. Daneyko, A. Höltzel, A. Seidel-Morgenstern, and U. Tallarek. Statistical analysis of packed beds, the origin of short-range disorder, and its impact on eddy dispersion. *Journal of Chromatography A*, 1217(28):4713–4722, 2010.

[101] T. H. Kjeldsen. A contextualized historical analysis of the kuhn–tucker theorem in nonlinear programming: the impact of world war II. *Hist. Math.*, 27(4):331–361, 2000.

[102] M. A. Klatt and S. Torquato. Characterization of maximally random jammed sphere packings. ii. correlation functions and density fluctuations. *Physical Review E*, 94(2):022152, 2016.

[103] H. Klüpfel. The simulation of crowd dynamics at very large events—calibration, empirical data, and validation. In *Pedestrian and Evacuation Dynamics 2005*, pages 285–296. Springer, 2007.

[104] A. N. Kolmogorov. Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Bull Univ État Moscou Sér Int A*, 1:1–26, 1937.

[105] M. H. Köpf and L. M. Pismen. A continuum model of epithelial spreading. *Soft Matter*, 9(14):3727–3734, 2013.

[106] J.-U. Kreft, G. Booth, and J. Wimpenny. BacSim, A simulator for individual-based modelling of bacterial colony growth. *Microbiology*, 144:3275–3287, 1998.

[107] J.-U. Kreft, G. Booth, and J. W. Wimpenny. BacSim, A simulator for individual-based modelling of bacterial colony growth. *Microbiology*, 144(12):3275–3287, 1998.

[108] G. Le Dréau, M. Saade, I. Gutiérrez-Vallejo, and E. Martí. The strength of smad1/5 activity determines the mode of stem cell division in the developing spinal cord. *J Cell Biol*, 204(4):591–605, 2014.

[109] H. O. Lee and C. Norden. Mechanisms controlling arrangements and movements of nuclei in pseudostratified epithelia. *Trends in cell biology*, 23(3):141–150, 2013.

[110] R. S. Lee and R. L. Hughes. Prediction of human crowd pressures. *Accident analysis & prevention*, 38(4):712–722, 2006.

[111] R. Leine and H. Nijmeijer. *Dynamics and Bifurcations of Non-Smooth Mechanical Systems*, volume 18. Springer Science & Business Media, 2013.

[112] M. Leroy-Lerêtre, G. Dimarco, M. Cazales, M.-L. Boizeau, B. Ducommun, V. Lobjois, and P. Degond. Are tumor cell lineages solely shaped by mechanical forces? *Bulletin of mathematical biology*, 79(10):2356–2393, 2017.

[113] J. Li. Macroscopic model for head-on binary droplet collisions in a gaseous medium. *Phys. Rev. Lett.*, 117:214502, 2016.

[114] Y. Li and T. Tanaka. Kinetics of swelling and shrinking of gels. *The Journal of chemical physics*, 92(2):1365–1371, 1990.

[115] S. Maheswaran, P. M. Speight, and P. Hammond. Modeling epithelial cell behavior and organization. *IEEE transactions on nanobioscience*, 6(1):77–85, 2007.

[116] S. N. Majumdar, K. Mallick, and S. Sabhapandit. Statistical properties of the final state in one-dimensional ballistic aggregation. *Physical Review E*, 79(2):021109, 2009.

[117] H. A. Makse, J. Brujić, and S. F. Edwards. Statistical mechanics of jammed matter. *Phys. Granul. Media*, pages 45–85, 2004.

[118] T. Mammoto and D. E. Ingber. Mechanical control of tissue and organ development. *Development*, 137(9):1407–1420, 2010.

[119] P. A. Martin and J. Piasecki. One-dimensional ballistic aggregation: Rigorous long-time estimates. *Journal of statistical physics*, 76(1-2):447–476, 1994.

[120] P. A. Martin and J. Piasecki. Aggregation dynamics in a self-gravitating one-dimensional gas. *Journal of statistical physics*, 84(3-4):837–857, 1996.

[121] B. Maury. Direct stimulation of aggregation phenomena. *Communications in Mathematical Sciences*, 2(Supplemental Issue):1–11, 2004.

[122] B. Maury. A time-stepping scheme for inelastic collisions. *Numer. Math.*, 102(4):649–679, 2006.

[123] B. Maury, A. Roudneff-Chupin, F. Santambrogio, and J. Venel. Handling congestion in crowd motion models. *Net. Het. Media*, 6(3):485–519, 2011.

[124] B. Maury and J. Venel. Handling of contacts in crowd motion simulations. *Traffic and Granular Flow'07*, pages 171–180, 2009.

[125] B. Mehlig, M. Wilkinson, K. Duncan, T. Weber, and M. Ljunggren. Aggregation of inertial particles in random flows. *Physical Review E*, 72(5):051104, 2005.

[126] L. Meinecke and M. Eriksson. Excluded volume effects in on-and off-lattice reaction–diffusion models. *IET systems biology*, 11(2):55–64, 2016.

[127] F. A. Meineke, C. S. Potten, and M. Loeffler. Cell migration and organization in the intestinal crypt using a lattice-free model. *Cell proliferation*, 34(4):253–266, 2001.

[128] P. Michael and R. Zoltan. Active zone of growing clusters: Diffusion-limited aggregation and the eden model. *Phys. Rev. Lett.*, 53:415, 1984.

[129] J. J. Moreau. Unilateral contact and dry friction in finite freedom dynamics. In *Nonsmooth mechanics and Applications*, pages 1–82. Springer, 1988.

[130] S. Motsch and D. Peurichard. From short-range repulsion to hele-shaw problem in a model of tumor growth. *Journal of mathematical biology*, 76(1-2):205–234, 2018.

[131] S. Müller-Pfeiffer, H.-J. Anklam, and W. Haubenreisser. A generalized ballistic aggregation model for the simulation of thin film growth with special consideration of nodular growth. *physica status solidi (b)*, 160(2):491–504, 1990.

[132] J. D. Murray. Mathematical biology. i, volume 17 of interdisciplinary applied mathematics, 2002.

[133] O. R. Musin. The kissing number in four dimensions. *Annals of Mathematics*, pages 1–32, 2008.

[134] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, USA, second edition, 2006.

[135] A. Okabe. *Spatial tessellations*. Wiley Online Library, 1992.

[136] S. Okuzumi, H. Tanaka, and M.-a. Sakagami. Numerical modeling of the coagulation and porosity evolution of dust aggregates. *The Astrophysical Journal*, 707(2):1247, 2009.

[137] G. Y. Onoda and E. G. Liniger. Random loose packings of uniform spheres and the dilatancy onset. *Phys. Rev. Lett.*, 64:2727–2730, May 1990.

[138] J. M. Osborne, A. G. Fletcher, J. M. Pitt-Francis, P. K. Maini, and D. J. Gavaghan. Comparing individual-based approaches to modelling the self-organization of multicellular tissues. *PLoS computational biology*, 13(2):e1005387, 2017.

[139] L. Paoli and M. Schatzman. A numerical scheme for impact problems i: The one-dimensional case. *SIAM Journal on Numerical Analysis*, 40(2):702–733, 2002.

[140] L. Paoli and M. Schatzman. A numerical scheme for impact problems ii: The multidimensional case. *SIAM journal on numerical analysis*, 40(2):734–768, 2002.

[141] J. K. Parrish and W. M. Hamner. *Animal groups in three dimensions: how species aggregate*. Cambridge University Press, 1997.

[142] S. N. Pathak, D. Das, and R. Rajesh. Inhomogeneous cooling of the rough granular gas in two dimensions. *EPL (Europhysics Letters)*, 107(4):44001, 2014.

[143] N. Pelechano and A. Malkawi. Evacuation simulation models: Challenges in modeling high rise building evacuation with cellular automata approaches. *Automation in Construction*, 17(4):377 – 385, 2008.

[144] M. Pelliccione and T.-M. Lu. *Evolution of Thin Film Morphology: Modelling and Simulations*. Springer-Verlag Berlin Heidelberg, 2008.

[145] B. Perthame, F. Quirós, M. Tang, and N. Vauchelet. Derivation of a Hele-Shaw type system from a cell model with active motion. *arXiv preprint arXiv:1401.2816*, 2014.

[146] B. Perthame, F. Quirós, and J. L. Vázquez. The Hele–Shaw asymptotics for mechanical models of tumor growth. *Arch. Ration. Mech. Anal.*, 212(1):93–127, 2014.

[147] B. Perthame and N. Vauchelet. Incompressible limit of a mechanical model of tumour growth with viscosity. *Phil. Trans. R. Soc. A*, 373(2050):20140283, 2015.

[148] D. W. Peterson. A review of constraint qualifications in finite-dimensional spaces. *SIAM Review*, 15(3):639–654, 1973.

[149] D. Peurichard, F. Delebecque, A. Lorsignol, C. Barreau, J. Rouquette, X. Descombes, L. Casteilla, and P. Degond. Simple mechanical cues could explain adipose tissue morphology. *Journal of theoretical biology*, 429:61–81, 2017.

[150] M. J. Plank and M. J. Simpson. Models of collective cell behaviour with crowding effects: comparing lattice-based and lattice-free approaches. *Journal of the Royal Society Interface*, 9(76):2983–2996, 2012.

[151] R. Poliquin and R. Rockafellar. Tilt stability of a local minimum. *SIAM Journal on Optimization*, 8(2):287–299, 1998.

[152] M. Rabatel. *Modélisation Mécanique des Glaces de Mer*. PhD thesis, Université de Grenoble, 2015.

[153] M. Rabatel, S. Labbé, and J. Weiss. *Dynamics of an Assembly of Rigid Ice Floes*. J. Geophys. Res. Oceans 120, no. 9, 5887 to 5909, 2015.

[154] P. Ramanlal and L. Sander. Theory of ballistic aggregation. *Physical review letters*, 54(16):1828, 1985.

[155] D. Rapaport. The event scheduling problem in molecular dynamic simulation. *Journal of Computational Physics*, 34(2):184–201, 1980.

[156] J. Roberts. *Sticky Pixels: Evolutionary Growth by Random Drop Ballistic Aggregation*. Research Gate, 2014.

[157] E. Ronchi and D. Nilsson. Fire evacuation in high-rise buildings: a review of human behaviour and modelling research. *Fire science reviews*, 2(1):7, 2013.

[158] T. Roose, S. J. Chapman, and P. K. Maini. Mathematical models of avascular tumor growth. *SIAM Rev.*, 49(2):179–208, 2007.

[159] S. A. Sandersius, M. Chuai, C. J. Weijer, and T. J. Newman. Correlating cell behavior with tissue topology in embryonic epithelia. *PLoS One*, 6(4):e18081, 2011.

[160] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

[161] G. Saussine, C. Cholet, P. Gautier, F. Dubois, C. Bohatier, and J.-J. Moreau. Modelling ballast behaviour under dynamic loading. part 1: A 2d polygonal discrete element method approach. *Computer methods in applied mechanics and engineering*, 195(19):2841–2859, 2006.

[162] G. Schaller and M. Meyer-Hermann. Continuum versus discrete model: a comparison for multicellular tumour spheroids. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 364(1843):1443–1464, 2006.

[163] M. Scianna, L. Preziosi, and K. Wolf. A cellular potts model simulating cell migration on and in matrix environments. *Mathematical Biosciences and Engineering*, 10(1):235–261, 2013.

[164] G. Scott and D. Kilgour. The density of random close packing of spheres. *Journal of Physics D: Applied Physics*, 2(6):863, 1969.

[165] I. Smart. Proliferative characteristics of the ependymal layer during the early development of the mouse diencephalon, as revealed by recording the number, location, and plane of cleavage of mitotic figures. *Journal of anatomy*, 113(Pt 1):109, 1972.

[166] C. Song, P. Wang, and H. A. Makse. A phase diagram for jammed matter. *Nature*, 453(7195):629, 2008.

[167] Y. Song, Z. Wang, L. Zhao, and J. Luo. Dynamic crushing behavior of 3d closed-cell foams based on voronoi random model. *Materials & Design*, 31(9):4281–4289, 2010.

[168] D. Staple, R. Farhadifar, J.-C. Röper, B. Aigouy, S. Eaton, and F. Jülicher. Mechanics and remodelling of cell packings in epithelia. *The European Physical Journal E*, 33(2):117–127, 2010.

[169] J. O. Stoner. Properties of evaporated target foils studied by computer simulation of ballistic aggregation. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 282(1):242–246, 1989.

[170] S. J. Streichan, C. R. Hoerner, T. Schneidt, D. Holzer, and L. Hufnagel. Spatial constraints control cell proliferation in tissues. *Proceedings of the National Academy of Sciences*, 111(15):5586–5591, 2014.

[171] J. L. Synge and A. Schild. *Tensor calculus*, volume 5. Courier Corporation, 1969.

[172] E. Theveneau and R. Mayor. Neural crest delamination and migration: from epithelium-to-mesenchyme transition to collective cell migration. *Developmental biology*, 366(1):34–54, 2012.

[173] S. Torquato, T. M. Truskett, and P. G. Debenedetti. Is random close packing of spheres well defined? *Physical review letters*, 84(10):2064, 2000.

[174] S. Troyanski. Gateaux differentiable norms in lp. *Mathematische Annalen*, 287(1):221–227, 1990.

[175] T. Vicsek. *Fractal Growth Phenomena*. World Scientific, 1992.

[176] D. Walker, B. Brown, D. Hose, and R. Smallwood. Modelling the electrical impedivity of normal and premalignant cervical tissue. *Electronics Letters*, 36(19):1603–1604, 2000.

[177] D. Walker, J. Southgate, G. Hill, M. Holcombe, D. Hose, S. Wood, S. Mac Neil, and R. Smallwood. The epitheliome: agent-based modelling of the social behaviour of cells. *Biosystems*, 76(1-3):89–100, 2004.

[178] E. W. Weisstein. Kepler conjecture. *Wolfram Research, Inc.*, 2003.

[179] S. Wolfram. *Cellular automata and complexity: collected papers*. CRC Press, 2018.

[180] W. Xu, H. Chen, and Z. Lv. A 2D elliptical model of random packing for aggregates in concrete. *J. of Wuhan Univ. Technol.-Mater. Sci. Ed.*, 25(4):717–720, 2010.

[181] R. Zallen. *The Physics of Amorphous Solids*. Wiley/Interscience, New York, 1983.

[182] J. Zimmermann, B. A. Camley, W.-J. Rappel, and H. Levine. Contact inhibition of locomotion determines cell–cell and cell–substrate forces in tissues. *Proceedings of the National Academy of Sciences*, 113(10):2660–2665, 2016.

[183] E. Zlotek-Zlotkiewicz, S. Monnier, G. Cappello, M. Le Berre, and M. Piel. Optical volume and mass measurements show that mammalian cells swell during mitosis. *J Cell Biol*, 211(4):765–774, 2015.

[184] W. Zulehner. Analysis of iterative methods for saddle point problems: a unified approach. *Mathematics of computation*, 71(238):479–505, 2002.